Linguistic Typology at the Crossroads 3-2 (2023): 132-190 ISSN 2785-0943

The Dimensions of Morphosyntactic Variation: Whorf, Greenberg and Nichols were right

SIVA KALYAN[1] & MARK DONOHUE[2]

[1]THE UNIVERSITY OF QUEENSLAND & THE AUSTRALIAN NATIONAL UNIVERSITY,

[2]THE LIVING TONGUES INSTITUTE FOR ENDANGERED LANGUAGES

Submitted: 27/06/2023 Revised version: 23/10/2023

Accepted: 27/11/2023 Published: 27/12/2023

Abstract

We examine a database of 3089 languages coded for 351 morphosyntactic features, including almost all of the morphosyntactic features found in The World Atlas of Language Structures (Dryer & Haspelmath 2013). We apply Factor Analysis of Mixed Data, and determine that the main dimensions of global morphological variation involve (1) word order in clauses and adpositional phrases, (2) head- versus dependent-marking, and (3) a set of features that show an east-west distribution. We find roughly the same features clustering in similar dimensions when we examine individual macro-areas, thus confirming the universal relevance of these groupings of features, as encapsulated in well-known implicational universals. This study confirms established insights in linguistic typology, extending earlier research to a much larger set of languages, and uncovers a number of areal patterns in the data.

Keywords: typology; word order; morphosyntax; head/dependent-marking, computational linguistics; areality.

1. Introduction

The goal of much early work in linguistic typology was to categorise languages into distinct overall “types”, under the assumption that once the type of a language was known, a large number of its features could then be predicted – in effect, a holistic

approach to typology (Croft 2003: 31, Humboldt 1836, von der Gabelentz 1901). While the prominence of such taxonomic work has receded in favour of detailed studies of individual features (or clusters of related features), the time is ripe to resuscitate such work in light of the increased amount of linguistic data that has become available. In this paper, we use Factor Analysis of Mixed Data (FAMD; Pagès 2004) to determine the main dimensions of global typological variation in morphosyntax – that is, the features that are most helpful for dividing the world's languages into different morphosyntactic types. We use a large database of morphosyntactic features, built and substantially expanded from the *WALS* dataset (Dryer & Haspelmath 2013; see Appendix 2), with both features and languages chosen independently of the present study. Upon examining the principal dimensions emerging from this analysis, we find that in each case, they bring together a group of features that has previously been proposed as a basis for the global typological classification of languages: in particular, we find groups of features relating to (mostly clausal) word order (proposed as a basis for classification by Greenberg 1963, refined by Dryer 1992, 2013, and other publications; Dimensions 1 and 4; sections 3.1.1 and 3.1.4), head/dependent marking (Nichols 1986; Dimensions 1 and 2; sections 3.1.1 and 3.1.2), and a set of features that define a global east-west split, one end of which is mainly present in the Old World, and the other end of which is dominant in the "Circum-Pacific" region (Bickel & Nichols 2006, Bugaeva et al. 2021; section 3.1.3).

The remainder of this paper is structured as follows. In section 2, we introduce our dataset (World_morphosyntax), the metadata used as controls, and the technique of FAMD. In section 3, we present our results, for the global set of languages as well as for each macro-area individually; we find that the groupings of features that emerge in the global analysis recur in individual macro-areas. In section 4 we examine some of the negative results, discussing the kinds of features that have the smallest contribution to the global analysis. Finally, in section 5 we summarise our findings, and suggest directions for future research. A number of appendices illustrate the distribution of the individual features that emerge as relevant to defining the four dimensions described in section 3.[1]

[1] Appendices are available as supplementary material at:
https://typologyatcrossroads.unibo.it/article/view/17482/17369

2. Data and methodology

The World_morphosyntax dataset consists of 3089 language varieties (rows—see Appendix 10), representing 2,693 distinct iso 639-3 codes,[2] coded for 351 morphosyntactic features (columns). It is curated by Mark Donohue (see Appendix 1), and has been developed since 2010. The original database was based on the most robustly coded languages and features from the *World Atlas of Language Structures* (Dryer & Haspelmath 2013). The dataset includes most of the 155 morphosyntactic features in *WALS*, with unordered multivalued features recoded as sets of binary features. (See Appendix 2 for a full listing and description of the features in the World_morphosyntax dataset.) For instance, *WALS* feature 57A ('Position of Pronominal Possessive Affixes') is coded as a single feature in *WALS*, consisting of the features listed in (1).

(1) *WALS* feature 57A 'Position of Pronominal Possessive Affixes'

a. Possessive prefixes
b. Possessive suffixes
c. Prefixes and suffixes
d. No possessive affixes

We have recoded this single, categorial, features into three binary features, and added an additional feature, as listed in (2). This recoding captures the variation in *WALS* feature 57A, in M72 and M73; the 'Prefixes and suffixes' values of *WALS* 57A is coded with positive values for both of M72 and M73, thus showing commonality with both prefixal languages and suffixal languages, which is not automatically extracted from the *WALS* coding. Positive values for M72 and M73 are unified by M71, which captures the commonality between prefixal and suffixal marking in that both do represent the coding of features of the possessor on the possessum. M70 adds in a typologically-attested variable that is not coded in *WALS*.

[2] The most doubled iso codes are cmn (Mandarin varieties), zlm (Malay varieties), adi (Tani languages), each of which has ten or more entries, at least some of which represent different languages by any normal criteria.

(2) Features M70 – M73

a.	M70	Possession: associative tone	+/–
b.	M71	Possessive affixes: any	+/–
c.	M72	Possessive affixes: prefixes	+/–
d.	M73	Possessive affixes: suffixes	+/–

In addition to recoding some of the *WALS* features, additional features were added. *WALS* feature 102A codes for the appearance of agreement for A or P arguments. We have added coding for an S argument, as well as a third argument (M238), to account for languages that allow a recipient or dative argument to appear indexed on the verb. Very rarely a fourth or fifth agreement position can be found, and these are also coded, as M243 and M244. Additionally, just as *WALS* codes the position of agreement affixes marking possession on nouns, as prefixal or suffixal, we add in coding for the position of agreement affixes on verbs, as shown in (3).

(3) Coding the position of verbal agreement

a.	M232	verb agreement_S prefix	+/–
b.	M233	verb agreement_S suffix	+/–
c.	M234	verb agreement_A prefix	+/–
d.	M235	verb agreement_A suffix	+/–
e.	M236	verb agreement_P prefix	+/–
f.	M237	verb agreement_P suffix	+/–
g.	M239	verb agreement_R/D prefix	+/–
h.	M240	verb agreement_R/D suffix	+/–
i.	M245	verb agreement_tone A	+/–
j.	M246	verb agreement_tone S	+/–
k.	M247	verb agreement_tone P	+/–

Other *WALS* features that were recoded in order to enhance their matching with a related feature in the database are the features devoted to morphological causatives. These were recoded in line with the features focusing on applicatives (which were also expanded). In *WALS* applicatives are coded for the kinds of bases that allow applicative extensions (intransitive or transitive bases), and the semantic role of the

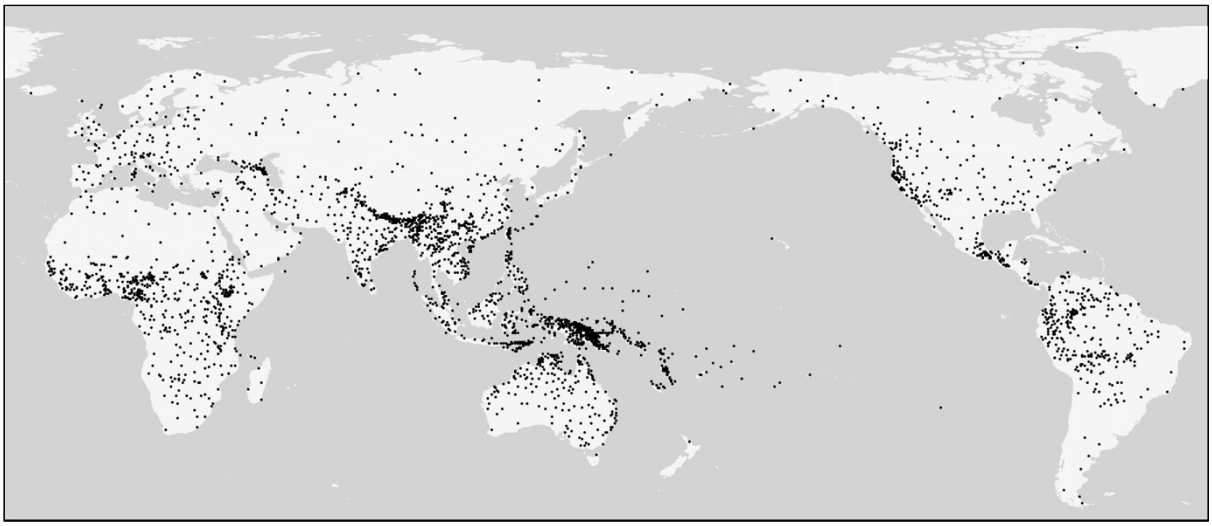
applied object (Benefactive, Instrumental, or Locative) (Polinsky 2013), and (non-periphrastic) causatives are coded according to whether they are morphological or compounds (Song 2013). In the World_morphosyntax database the range of semantic roles for applicatives was expanded (Benefactive, Instrumental, Locative, Associative/Comitative, Theme, Reason, and Malefactive), and the types of bases allowed were extended to query ditransitive bases, and to distinguish whether agentive or patientive intransitive bases (or both) permit applicatives. Additionally, the possibility of more than one applicative appearing on a single base was coded with two features, as well as the possibility that the 'applicative' construction promotes directly to subject (a 'superapplicative', as attested in many languages of Taiwan and the Philippines). Matching the detail on applicative constructions present in *WALS* and expanded on in World_morphosyntax, we coded morphological causatives according to whether they are attested in patientive or agentive intransitive bases, or even on ditransitive bases, as well as whether double (or second) applicatives are attested with different bases, what the coding strategy is for the causee of a causative construction with three arguments, and whether there is syncretism between the morpheme used for causatives and applicatives.

Other added features logically extend the scope of *WALS* features (for instance, explicitly coding more semantic roles that can be introduced with applicative constructions, the existence of suppletive negative verbal stems, or morphological processes other than prefixation and suffixation - namely, infixation, and metathesis). Wholly new features centre around the possibilities for nominal incorporation into verbs. Further details on the features in the database can be found in Appendix 2.

On average, the coding of languages reported here from the World_morphosyntax dataset is 86% complete; this compares favourably with *WALS* (18% of 155 features for 2662 languages), as well as more recent datasets such as Grambank (Skirgård et al. 2023; 76% of 195 features for 2430 languages); see Appendix 1. For this study, we excluded known pidgins and creoles, reconstructed proto-languages, and ancient or historical languages. Pidgins and creoles frequently represent lineages that are not original to the area in which they are currently found, and in most cases represent disruptions to the local typological landscape. Ancient or historical languages (i.e. those that are attested only before the era of European colonisation) are by definition not part of any modern linguistic ecology, and so should not be included in an analysis of modern languages. Excluding these, we were left with 3089 languages/varieties,

the locations of which are shown in Map 1.[3] (See Appendix 1 for a full listing of the languages, with their genealogical and areal memberships.)

To analyse the World_morphosyntax dataset, we used Factor Analysis of Mixed Data (FAMD; Pagès 2004), a dimensionality reduction technique that combines Multiple Correspondence Analysis (MCA) and Principal Components Analysis (PCA), as implemented in the FactoMineR package for R (Lê et al. 2008; see the Supplementary Materials for our annotated source code). Since MCA is suited for data consisting exclusively of categorical variables, and PCA is suited for data consisting exclusively of continuous variables, we felt that FAMD is the appropriate choice for the World_morphosyntax dataset, which contains 27 ordinal variables (which we treated as continuous) and 324 binary variables. We started by imputing[4] the missing values in the dataset using the regularised iterative FAMD algorithm, as implemented in the "imputeFAMD" function in the missMDA package for R (Josse & Husson 2016), using 4 components (though the number of components made little difference to the results).

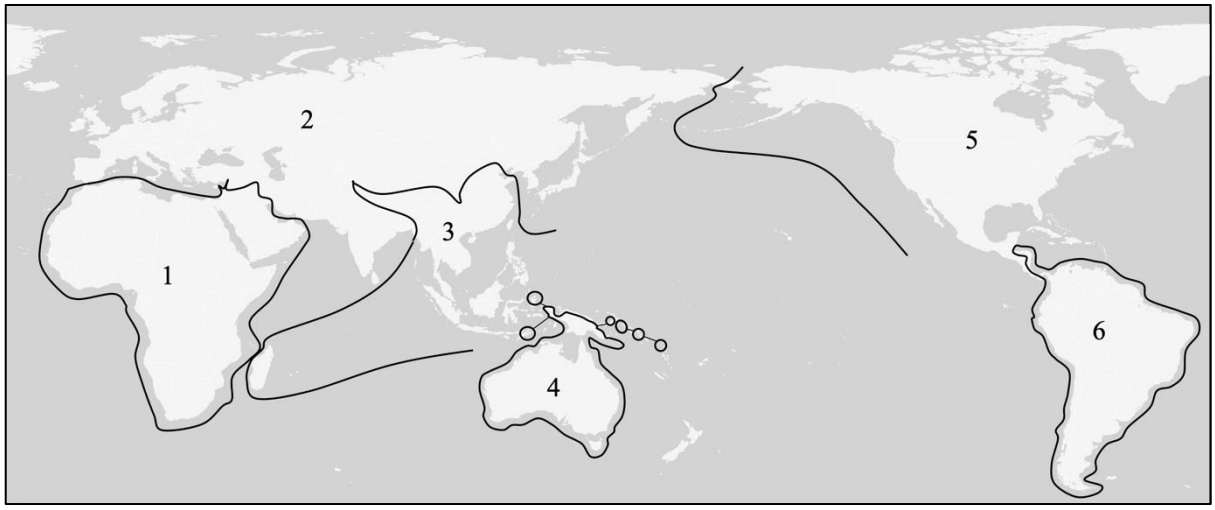


Map 1: Languages and language varieties included in the analysis ($n = 3089$).

[3] Legend: The map (as well as subsequent maps) shows the world from 60º S to 90º N, and from 30º west extending 360º to the east.

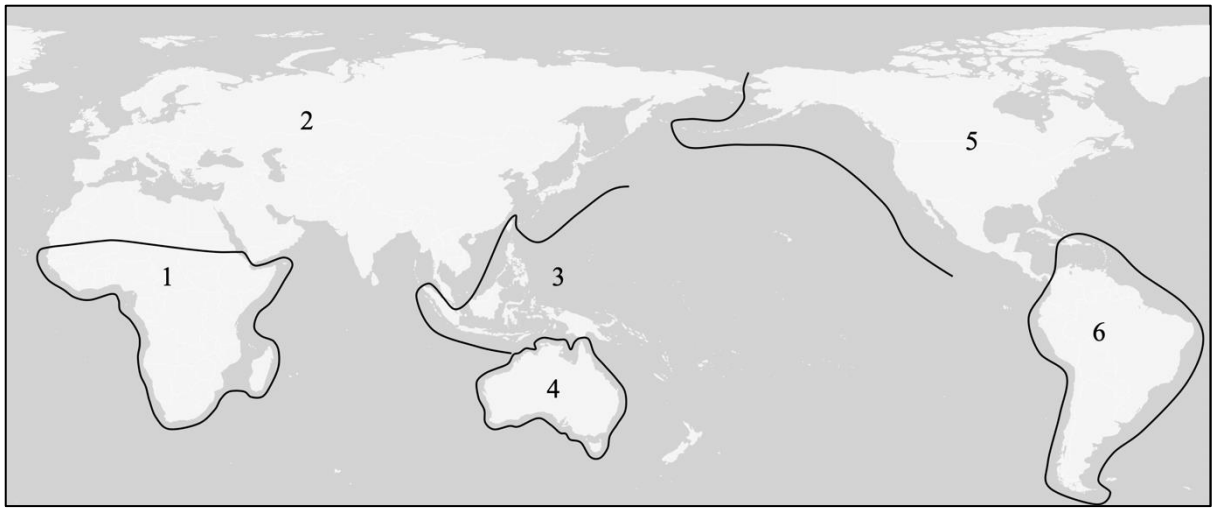
[4] Imputation is a family of techniques for replacing missing values in a dataset with estimates of the most likely values of those data points. It is necessary to perform imputation when applying techniques such as FAMD, since such techniques involve computing a covariance matrix, which requires complete data. The iterative FAMD algorithm for imputation (which we use here) works by first replacing missing values in each column with the column mean; then performing FAMD; then reconstructing the missing values based on the FAMD result; then performing FAMD again; and so on until the imputed values stabilise.

We then applied FAMD to the imputed data, assigning weights to languages in a way that equalises the total weights of different macro-areas, as well as of the different AUTOTYP areas within each macro-area; this was done to increase the likelihood that the dimensions that we find would capture groupings of features that are valid across different macro-areas, and across different areas within each macro-area. Macro-areas were defined according to Hammarström & Donohue (2014), itself a refinement of the macro-areas established by Dryer (1989, 1992), with the exception that languages belonging to the AUTOTYP "North Africa" area were re-assigned from Africa to Eurasia, on the grounds that whereas the Sahara Desert has been a barrier to contact since the end of the African Humid Period (e.g., de Menocal et al. 2000), cross-Mediterranean societies have flourished since antiquity. The two different macro-area divisions are compared in Maps 2 and 3. Most of the changes involve the abandonment of the apparent principle of unifying families into single macro-areas, and the split of Australia and (some of) New Guinea into separate macro-areas.



Map 2: Six macro-areas, following Dryer (1989, 1992).

Legend: 1: Africa; 2: Eurasia; 3: Southeast Asia and Oceania; 4: Australia-New Guinea; 5: North America; 6: South America.

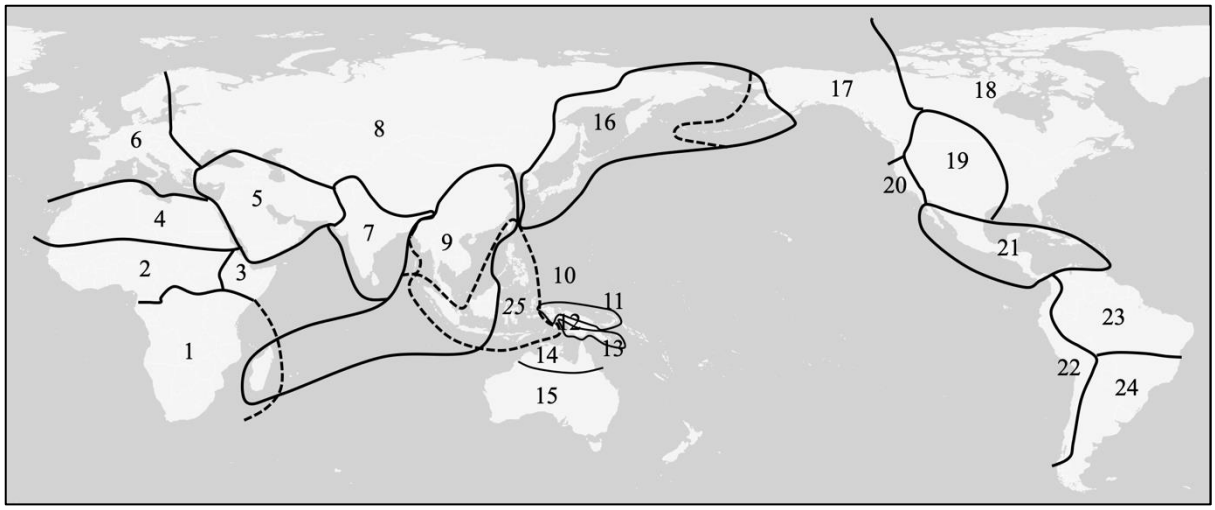


Map 3: Six macro-areas, following Hammarström and Donohue (2014), amended here.

Legend: 1: Africa; 2: Eurasia; 3: Pacific; 4: Australia; 5: North America; 6: South America.

The AUTOTYP areas were extrapolated from those described in Bickel et al. (2023), following Bickel (2002) and Nichols et al. (2013) (also https://www.autotyp.uzh.ch), with a few major differences: (1) 'Southeast Asia' has been split into Mainland Southeast Asia (consisting of the Southeast Asian languages of mainland Eurasia and Hainan) and Island Southeast Asia (consisting of the remaining Southeast Asian languages, as well as languages from 'Oceania' west of New Guinea and up to Taiwan), and Madagascar has been moved to Africa, allowing the smaller areas to be embedded unproblematically into macro-areas (as shown in Table 1); (2) The Andaman islands are grouped with "Indic", based on historical connections; (3) The languages of the Aleutian Islands are included in 'Alaska-Oregon', rather than 'North Coast Asia', based on geography and cultural connections. The different areas are shown in Map 4, contrasting the original Autotyp areas with the modified set used here.[5] Details of the assignment of individual languages to areas can be found in Appendix 1.

[5] Legend for Map 4: 1: S Africa; 2: African Savannah; 3: Greater Abyssinia; 4: N Africa; 5: Greater Mesopotamia; 6: Europe; 7: Indic; 8: Inner Asia; 9: Southeast Asia (mainland); 10: Oceania; 11: N Coast New Guinea; 12: Interior New Guinea; 13: S New Guinea; 14: N Australia; 15: S Australia; 16: N Coast Asia; 17: Alaska-Oregon; 18: E North America; 19: Basin and Plains; 20: California; 21: Mesoamerica; 22: Andean; 23: NE South America; 24: SE South America; 25: Island Southeast Asia.



Map 4: The 25 modified AUTOTYP areas compared to the original 24 areas.

We are aware of alternative ways of controlling for area and genealogy (e.g. Guzmán-Naranjo & Becker 2021 on phylogenetic regression and Gaussian processes; Macklin-Cordes & Round 2022 on phylogenetic weighting). However, we opted to stay with areally-weighted FAMD, for the sake of simplicity, and because the patterns we find are strong enough to be visible and consistent regardless of what controls we use (see Appendix 4).

Macro-area (6)	**Modified AUTOTYP area (25)**	***n* (languages)**
Africa	Africa, African Savannah, Greater Abyssinia	535
Eurasia	N Africa, Greater Mesopotamia, Europe, Inner Asia, Southeast Asia (mainland), N Coast Asia	1024
Pacific	Island Southeast Asia, N Coast New Guinea, Interior New Guinea, S New Guinea, Oceania	760
Australia	N Australia, S Australia	205
North America	Alaska-Oregon, E North America, Basin and Plains, California, Mesoamerica	283
South America	Andean, NE South America, SE South America	282

Table 1: Modified AUTOTYP areas arranged by Macro-area.

Another advantage of using areal divisions as a control is that the difference in size between the smallest group and the largest group is less than the difference between the size of the smallest language family or genus (namely 1) and the largest. This means that the area-based controls do not give undue weight to isolates and singleton

genera. An additional advantage of using areas, rather than genealogies, is that we avoid having to make decisions about controversial language families like Nilo-Saharan (Dimmendaal 2011), Trans-New Guinea (Pawley & Hammarström 2018), Transeurasian/Altaic (Clauson 1956, Schönig 2003), Austric (Schmidt 1906, Reid 2005), Hokan and Penutian (Campbell 1997, DeLancey & Golla 1997, Poser 1995), or Dene-Yeniseian (Kari & Potter 2010), or subgroups within families (e.g., Indo-Iranian and the position of Nuristani languages within Indo-European, the existence of Italo-Celtic in the same family, or the internal hierarchy of Tibeto-Burman). A comparison of the results presented here and the (minimally different) results of using genealogically-weighted approaches are discussed in Appendix 4.

3. Results

We examine the results in detail for the world as a whole, and then in summary for each of six macro-areas. Section 3.1 presents the global dimensions of variation, what linguistic features characterise these dimensions, and where languages displaying the highs and lows of these dimensions can be found.[6] In Section 3.2 we examine the dimension plots presented in Figure 3 to show where various areal or genealogical entities can be found, and to what degree they form 'compact' clusters in typological space. In Section 3.3 we examine whether, and to what extent, these feature groupings can be considered universal, based on their appearance in the separate analyses of individual macro-areas.

3.1. Overall

Figure 1 shows the percentage of total variance explained by each of the first 20 dimensions of the FAMD result, with the 'elbow' indicated by the arrow. As we can see, there is a sharp drop-off on the scree plot after the first four dimensions; thus, following principles in Cattell (1966), in the following we consider only the first four dimensions.

[6] We present four dimensions of variation, for the reasons discussed in Section 2.

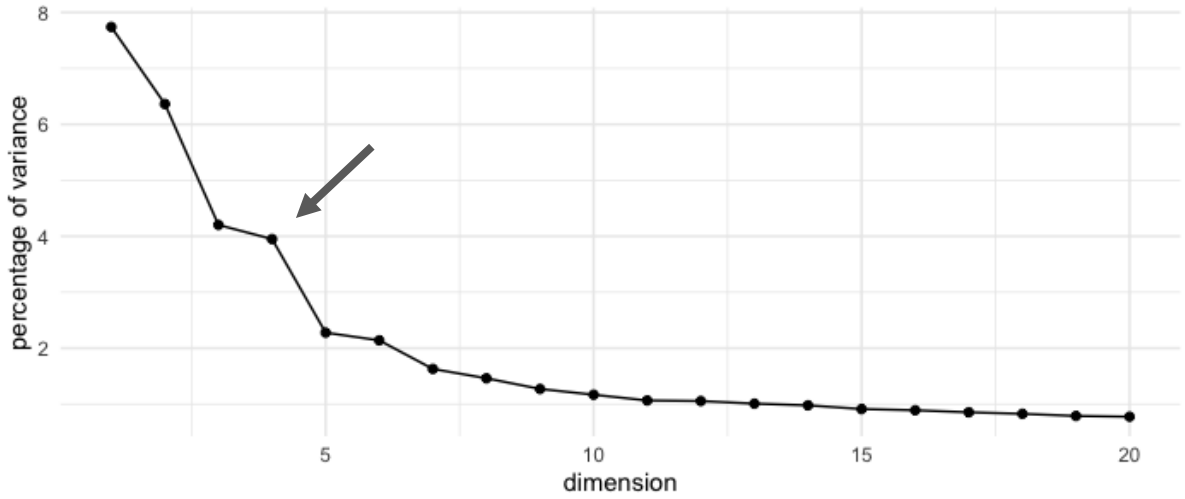


Figure 1: Scree plot showing variance accounted for by the first 20 dimensions.

The positions of languages according to these four dimensions are plotted in Figure 3. The leftmost column shows Dimension 1 along the *x* axis, and Dimensions 2, 3 and 4 on the *y* axis in rows 2, 3 and 4, respectively. In the second column Dimension 2 is shown on the *x* axis, and Dimension 1 is displayed on the *y* axis. In the third column Dimension 3 is plotted on the *x* axis, and in the fourth column Dimension 4 is plotted on the *x* axis, with the *y* axes representing the same dimensions as previously described. The colours of the dots vary according to their positions on the first, second and third dimensions, with these dimensions mapped to red, green and blue colour components, respectively (a technique exemplified in Nerbonne 2009, and other associated works). Combinations of red and green display as yellow, red + blue as purple/magenta, red + green + blue as white. Green + blue is cyan, and the absence of any colouring is black, as shown schematically in Figure 2 (dots can of course also occupy positions inside the cube, where the display colour tends towards grey). Note that in Figure 3 (and later in Map 12) Dimension 4 is not represented in the colours displayed (though see Appendix 8). These four dimensions in total account for 22.2% of the variance in the data (a figure comparable to, for example, Skirgård et al. 2023), as shown in Table 2.[7]

[7] Much of the remaining data can be divided into a) rare features; b) wide-spread common features without strong correlations with other grammatical features; c) geographically restricted features. This is discussed in Section 4. Section 3.3 examines the contribution of other features in determining variation in smaller regions (see also Appendix 5).

Dimension	Variance accounted for?	Section
1	7.7%	3.1.1
2	6.4%	3.1.2
3	4.2%	3.1.3
4	3.9%	3.1.4

Table 2: Variance in the data accounted for by the first four dimensions.

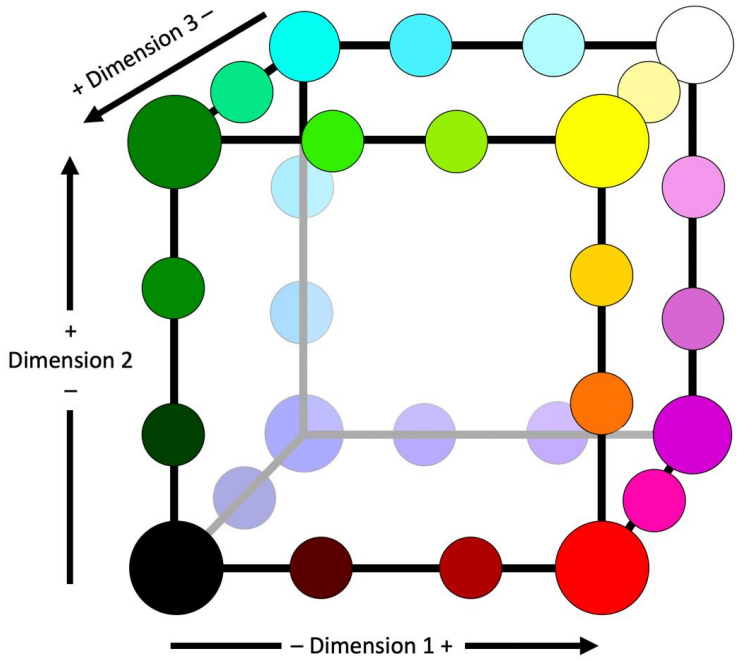


Figure 2: An illustration of a 'Red-Green-Blue' cube.

The interpretation of the different dimensions is presented in 3.1.1 – 3.1.4; in summary, the top end of Dimension 1, shown in red and orange, indicates languages with prepositions, and a tendency towards subject prefixes on verbs, while the bottom end is occupied by SOV languages with case-marking, shown in green and blue. The top end of Dimension 2 correlates with morphologically elaborate verbs, marked in pale green, while the bottom end tends towards isolating languages, with magenta colours. Dimension 3 has languages with gender systems and plural marking on nouns at the top end; languages at this pole are generally brown, while the lower end of this dimension correlates with VOS order and clusivity contrasts, presenting in a mix of colours in Figure 3. The top of Dimension 4 correlates with VSO languages that have prenominal modifiers in the NP, while the lower end correlates with SV order, object prefixes on verbs, and noun-numeral orders. As noted above, the position of a language on Dimension 4 is not indicated by any particular colours in Figure 3, but

Figures A8.1, A8.2 and A8.3, as well as Figures A8.6, A8.8 and A8.9 in Appendix 8 show the effects of having Dimension 4 contributing to the colouring.

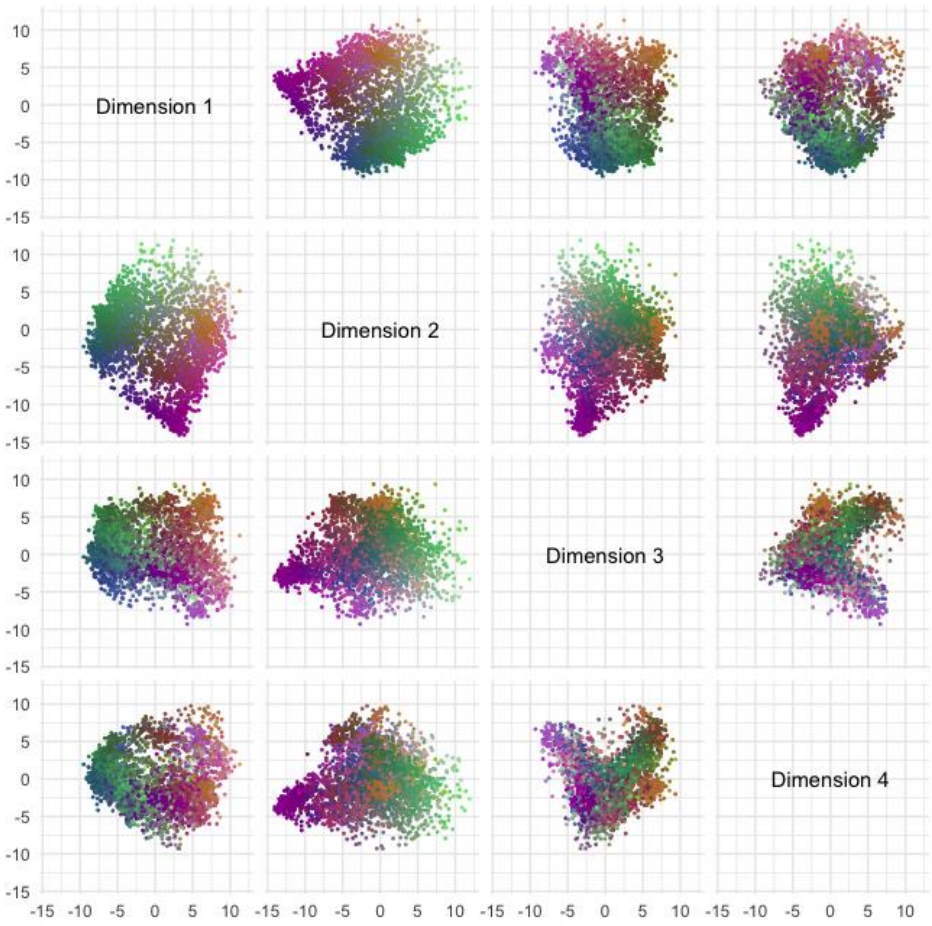


Figure 3: Languages plotted according to the first four dimensions of variance.

We can see that there are different 'densities' of languages in different areas in Figure 3, such as the paucity of languages at approximately (0, 3) in the plot of Dimension 1 vs. Dimension 4 (at the bottom left of Figure 3), and the high concentration of languages at (–6, 2.5) in the plot of Dimension 1 vs. Dimension 2 (at the top left of Figure 3).[8] The dimension plots, based purely on linguistic features, include a number of typologically differentiated or isolated regions that correspond with a high degree of precision to geographically-recognisable areas or genealogically-coherent entities, some of which are discussed below in 3.2 (and see Section 4 for further discussion).

[8] The low-density region corresponds to a mix of languages, including many from modern Iran such as Farsi (pes; west2369, Indo-European, Iranian) and Sorani (ckb; cora1257, Indo-European, Iranian), and the high-density region is occupied by the head-final languages displaying and extreme head-final typology such as is found in Turkish, Daghestanian, and other languages from central Eurasia.

In the following subsections, each dimension is characterised in terms of the features that show the strongest association with it; in the case of binary variables, the strength of this association is measured with an ANOVA test, and for continuous variables, it is measured using Pearson correlation. In both cases, we report an r^2 value. To determine whether the association is positive or negative, we look at the sign of the correlation coefficient (for continuous variables), or (for binary variables) perform a *t*-test comparing the dimension values of languages either exhibiting or lacking the feature against the entire set of languages, and note which (if either) of the two *t*-tests shows a significant positive value, and which (if either) shows a significant negative value. In Maps 5 – 8 positive values are shown in red/brown, and negative values in blue, according to the scale in Figure 4 (exact values can be found in Appendix 1).

Figure 4: The scale used in Maps 5 – 8.

3.1.1 Dimension 1: order of object and verb

The features most strongly associated with the first dimension centre around the order of the verb and its object, as well as a number of further headedness relations such as the position of a marker of subordination, the presence of prepositions, and the presence and position of case markers. Table 3 shows the features that have the strongest associations with Dimension 1.[9]

[9] For display purposes a number of related features from our database have been merged in this and subsequent tables for simplicity of presentation. For instance, both 'SOV' and 'OV' are associated (negatively) with Dimension 1 (since languages with these features on average have a negative value along Dimension 1; r^2 = 0.53 and 0.54, respectively). They are reported in Table 3 simply as SOV. Similarly, 'Core case (any)', 'Dependent marking', 'Number of cases' and 'Postnominal case' are all associated (negatively) with Dimension 1 (r^2 = 0.43, 0.49. 0.49 and 0.50, respectively), but only two of these features are listed in Table 3. Fuller lists of r^2 values are found in Appendix 1.

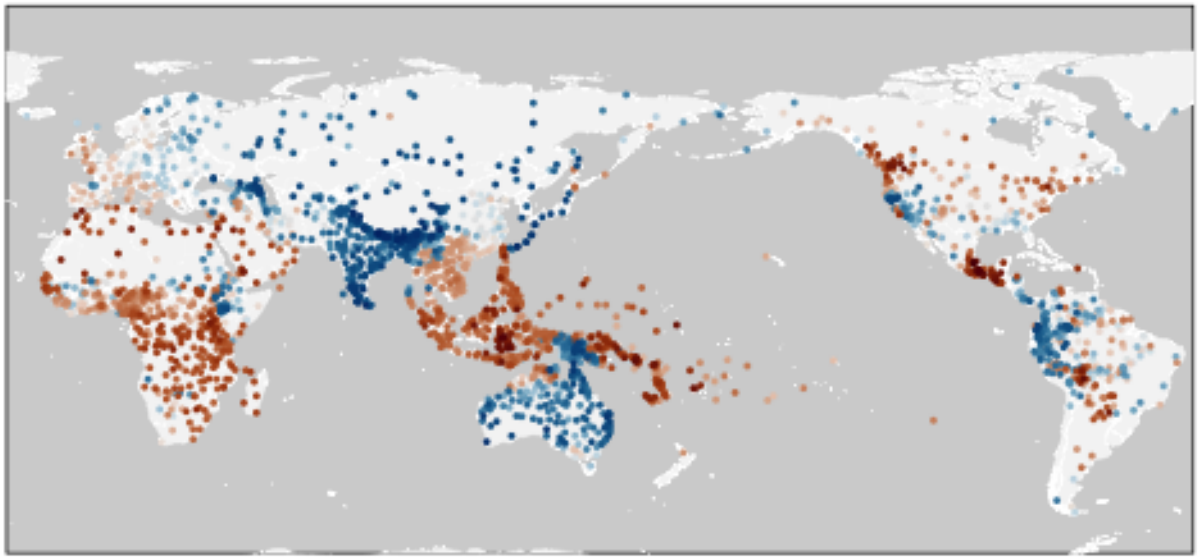
Direction	Feature	r^2
High	Prepositions	0.50
	Verb-Object order	0.42
	Initial subordination	0.41
	Nominative agreement by prefix	0.32
	Obliques follow verb	0.31
	Genitive precedes noun	0.31
	Final subordination by suffix	0.39
	Postpositions	0.42
	Obliques precede verb	0.45
	Number of cases	0.49
	Postnominal case	0.50
Low	SOV order	0.53

Table 3: Features characterising the extremities of Dimension 1.

These features are strongly reminiscent of (elements of) Greenberg's (1963) discussion of word order universals, and other linguistic features that refer to headedness parameters at the clause level. It is notable that prepositions are more closely associated with the positive (VO) end of this dimension than postpositions are with its negative (OV) end, and that head-final languages are strongly associated with the OV end, while head-initial languages are not as firmly associated with the VO end.

In Map 5 we can see the languages in our sample coded according to their position on Dimension 1, with high values marked in red/brown, and low values in blue, and middling values showing little hue. There are clear areal trends in the distribution of the extremes of this dimension, with large swathes of Eurasia dominated by languages with low (OV-congruent) values, and most of sub-Saharan Africa and Island Southeast Asia showing high (VO-congruent) values. Areas without consistent headedness settings, such as most of western Europe or northern China, are not associated with either extreme. The languages that are highest on Dimension 1 include various Otomanguean languages of the Chinantecan, Zapotecan and Popolocan groups in Central America, as well as Celebic Austronesian languages from central Indonesia, such as Mori (xmz; mori1268), Wolio (wlo; woli1241) and Wotu (wtw; wotu1240). The low end is dominated by South Asian languages, particularly South Dravidian

(Tamil, Tulu and Toda)[10] from the south of the subcontinent, and Bodic Tibeto-Burman (Kurtöp, Ghale and Balti)[11] from the Himalayas.



Map 5: Position of languages on Dimension 1 (blue = low, brown = high).

3.1.2 Dimension 2: verbal elaboration

The 2nd dimension of variation concerns the amount of morphology that can appear on the verb. At one end we have verbs with multiple positions for agreement, valency-increasing morphology for Ps (and, to a lesser extent, As), noun incorporation, and other inflectional material, such as switch-reference marking, Tense/Aspect/Mood, evidentiality, pluractionality, polarity, honorificity, voice marking, etc. (Bickel and Nichols 2013). At the other end, we find languages that lack extensive verbal morphology. The features with the strongest associations involve the lack of subordinating characteristics in "subordinate" clauses of different types, but the absence of the features characteristic of the higher end of this dimension, as well as the tendency for languages low on Dimension 2 to correlate with Dimension 1 (see Figure 3, and see 3.4), means that these languages tend to be more isolating.

[10] Tamil (tam; tami1289); Tulu (tcy; tulu1258); Toda (tcx; toda1252).

[11] Kurtöp (xkz; kurt1248); Ghale (ghe; barp1238); Balti (bft; balt1258).

Direction	Feature	r^2
High	Total verbal agreement positions	0.37
	Total verbal inflectional synthesis	0.34
	Total Modality affixes	0.30
	Incorporation	0.22
	Applicatives	0.20
	Causatives	0.16
	Possessive prefixes on nouns	0.15
	Total tense distinctions	0.11
	SVO order	0.14
	Symmetrical clauses: Purpose	0.17
	Symmetrical clauses: Temporal	0.18
Low	Symmetrical clauses: Reason	0.21

Table 4: Features characterising the extremities of Dimension 2.

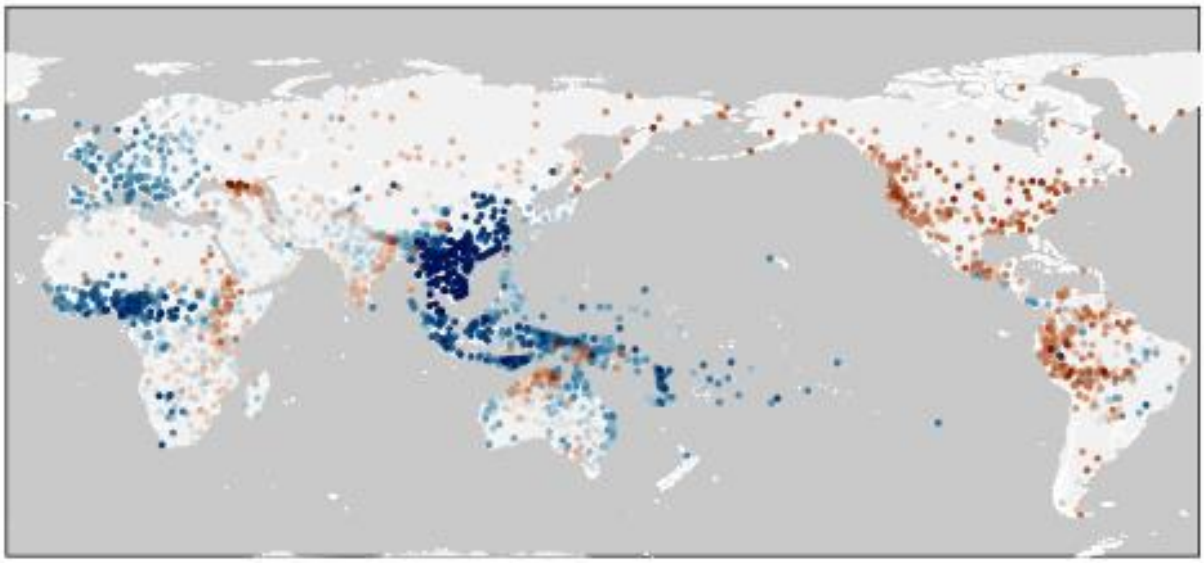
The features here are strongly reminiscent of (and add to) the head-marking end of Nichols' (1986) typology of head-marking vs. dependent-marking languages (itself related to divisions of morphological typology established as early as K.F. Schlegel 1808 and A.W. Schlegel 1818), with languages high on Dimension 2 being more heavily head-marking, and languages low on Dimension 2 showing more isolating/analytic traits. (We have already seen that dependent-marking is associated with Dimension 1, specifically with its lower OV end.)

In Map 6 we can see that the languages of the Americas are almost universally on the head-marking side of this dimension; the opposite extreme, namely absence of head-marking characteristics, dominates in Southeast Asia, to a lesser extent in western Africa, and in small measure in western Europe. The Old World sees clusters of head-marking languages in the Caucasus, in East Africa, in the Munda-Kiranti areas of South Asia, in the north-east of Eurasia on the approach to the Americas; parts of New Guinea, and most of northern Australia, also contain languages that are strongly head-marking, and so high on Dimension 2. Languages from a number of families in Southeast Asia are found at the low end of Dimension 2, including Austronesian (Moken, Cham)[12], Austroasiatic (Bruu, Vietnamese)[13], and also Hmong and Thai languages; a number of languages of West Africa, centred on Nigeria (such as Igede

[12] Moken (mwt; moke1242, Malayo-Sumbawan); Cham (cjm; east2563, Malayo-Sumbawan).

[13] Bruu (bru; east2332, Katuic); Vietnamese (vie; viet1252, Vietic).

and Yoruba)[14], are also high on this dimension. The higher end of the scale is occupied by polysynthetic languages from North America (such as Algonquian Arapaho, Cheyenne and Ottawa)[15], from the North-west Caucasus family (including Abaza, Adyghe and Kabardian)[16], as well as Chukotko-Kamchatkan Alyutor (alr; alut1245), and a number of western Amazonian languages from the South American (such as Aikanã, Jebero, Matses, Arakmbut and Matsigenka)[17], and a scattering of languages elsewhere.



Map 6: Position of languages on Dimension 2 (blue = low, brown = high).

3.1.3 Dimension 3: Western Old World

The third dimension of variation shows the strongest (macro-)areal distribution. Unlike the other three dimensions discussed here, the distribution of Dimension 3 does not identify a number of separate areas throughout the world, but rather a global cline from west to east (as is clearly visible in Map 7, and see below). The features at the high end of this dimension are all morphological; at the low end we see either an absence of extensive nominal morphology, or verb-initial orders. Because of these two

[14] Igede (ige; iged1239, Niger-Kongon, Idomoid); Yoruba (yor; yoru1245, Niger-Kongo, Yoruboid).

[15] Algonquian Arapaho (arp; arap1274); Cheyenne (chy; chey1247); Ottawa (otw; otta1242).

[16] Abaza (abq; abaz1241); Adyghe (ady; adyg1241); Kabardian (kbd; kaba1278).

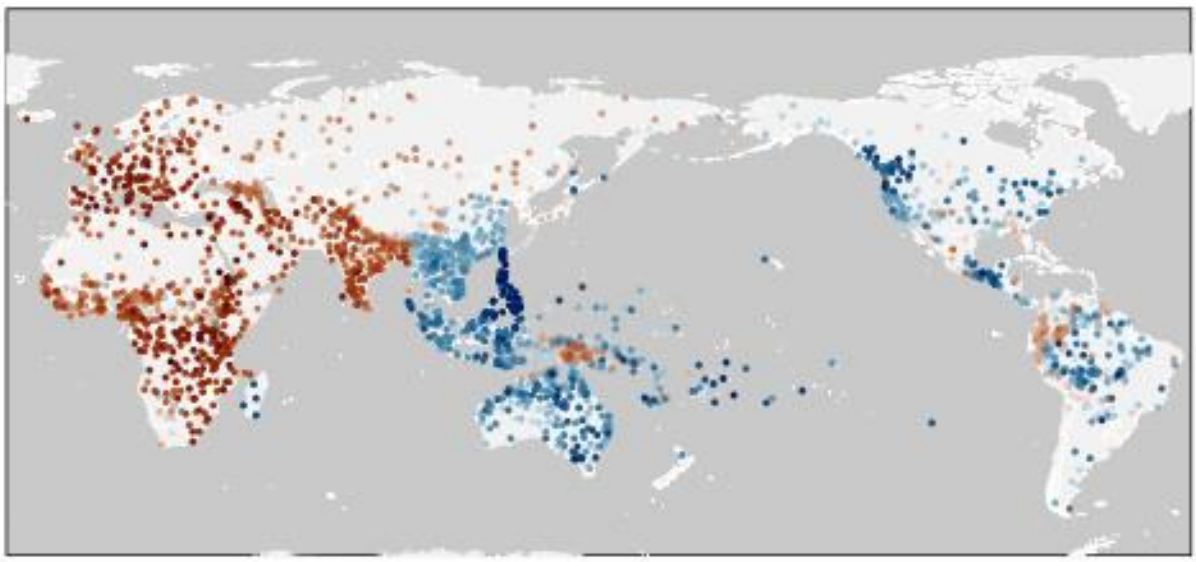
[17] Aikanã (tba; aika1237, Isolate, Aikanån); Jebero (jeb; jebe1250, Cahuapanan); Matses (mcf; mats1244, Panoan, Matses); Arakmbut (hug; arak1258, Harakmbet); Matsigenka (mcb, mach1267, Arawak, Campa).

different typologies, the r^2 values of features at the low end of this dimension are not as high as those at the high end.

Direction	Feature	r^2
High	Gender	0.28
	Obligatory plural marking on nouns	0.27
	3SG pronominal gender	0.25
	Verb alignment: accusative	0.22
	3PL pronominal gender	0.17
	Suffixal subject agreement on verbs	0.13
	Relative pronouns	0.13
	Ergativity	0.13
	VOS order	0.15
	Inclusive/Exclusive contrasts	0.18
Low	Clause-initial negation	0.18

Table 5: Features characterising the extremities of Dimension 3.

As can be seen in Map 7, languages low on this dimension are almost exclusively found in the Circum-Pacific region, an area which "comprises all of the Americas, Oceania (including Australia and New Guinea), and the mainland Asian Pacific Rim", the last area being the "Pacific-facing coast up to the lower slope of the far side of the major coast range" (i.e., the eastern Himalayas) (Bickel & Nichols 2006: 6). We observe increasingly high values as we go west in the Old World. On the high end we find most of the languages of Europe and the other circum-Mediterranean languages, as well as the Bantu languages, which have significantly higher values than those in the rest of Eurasia and Africa (see 3.2.1 and 3.2.10). In part due to this position, and the relative morphological simplicity of European languages compared to Semitic, Berber or Bantu languages (thus having lower values on Dimension 2), the languages of Europe can be identified as a global outlier (see Figure 3, and 3.2.1).



Map 7: Position of languages on Dimension 3 (blue = low, brown = high).

The features that have a strong association with Dimension 3 partially overlap with the list of features often put forward as defining 'Standard Average European' (Whorf 1941, and also Haspelmath 2001, van der Auwera 2011, and others). Often-cited 'Standard Average European' features that have positive correlations with Dimension 3 include: indefinite articles, have-perfects, relative pronouns (see Table 5), predominantly suffixing morphology (see Table 5), accusative alignment (see Table 5), and negative indefinite pronouns. Features that have negative associations with Dimension 3 include clusivity contrasts (see Table 5), alienability contrasts, identity of 'and' and 'with', and productive reduplication. In contrast to other studies on Standard Average European, we find that dative subjects have a (weak) positive association with Dimension 3 (r^2 = 0.10). (See Appendix 1 for details of the associations of these features with the different dimensions.)

It is clear from Map 7 that Dimension 3 negatively correlates with 'eastness' (as displayed on Map 7, such that Iceland is west and Greenland is east). The correlations of the different dimensions with 'eastness' in different domains are shown in Table 6. The strong negative correlation of Dimension 4 (3.1.4) with eastness in Eurasia reflects the far western position of the verb-initial Celtic, Berber and Semitic languages; there are very few verb-initial languages in the east of mainland Eurasia.

Dimension 3, however, shows strong correlations across Eurasia, the Old World, and globally.[18]

Dimension	Global	Old World	Eurasia
1	-0.02	-0.11	-0.45
2	0.36	-0.14	-0.22
3	-0.54	-0.61	-0.74
4	-0.21	-0.26	-0.73

Table 6: Correlation with eastness (*r*).

The languages at the top end of Dimension 3 are Niger-Kongo Bantu (Ruwund, KinyaRwanda and Runyankore)[19], Indo-European Romance (Spanish, Romansch, Galician and French)[20] or Afro-Asiatic Semitic (Cypriot Arabic, Mlaḥsô and Fezzan Arabic)[21], in addition to a number of other European languages (such as Albanian, Czech and Tabarchino)[22]. As can be seen in Figure 3, the lower end of Dimension 3 is quite dispersed typologically, and consequently there is a range of different languages that are maximally different from those of the western Old World, as measured on this dimension. Languages at the bottom of Dimension 3 include verb-initial Texistepec (poq; texi1237, Totozoquean, Chitimacha–Zoque), Kuikuro (kui; kuik1246, Cariban, Nahukwa), Shuswap (shs; shus1248, Salishan, Interior Salish), and many languages of the Philippines and Taiwan (such as Hanunoo, Saaroa and Maranao)[23], and Polynesia (Samoan and Niuean)[24] in the Pacific. In addition to their verb-initial clauses, these languages also lack gender in nouns or pronouns, accusative alignment, or obligatory plural marking.

[18] Strong negative correlations are also found in South America (-0.41), due to the presence of a large area in the northern Andean region occupied by languages with higher values on Dimension 3, belonging to the Jivaroan, Quechuan, Tucanoan and Boran families, amongst others.

[19] Ruwund (rnd; ruun1238); KinyaRwanda (kin; kiny1244); Runyankore (nyn; nyan1307).

[20] Spanish (spa; stan1288), Romansh (roh; roma1326), Galician (glg; gali1258), French (fra; stan1290).

[21] Cypriot Arabic (acy; cypr1248); Mlaḥsô (lhs; mlah1239); Fezzan Arabic (ayl; liby1240).

[22] Albanian (als; tosk1239, Indo-European, Albanian), Czech (ces; czec1258, Indo-European, Slavic); Tabarchino (lij; ligu1248, Indo-European, Romance).

[23] Hanunoo (hnn; hanu1241, Austronesian, Philippines); Saaroa (sxr; saar1237, Austronesian, Tsouic); Maranao (mrw; mara1404, Austronesian, Philippines).

[24] Samoan (smo; samo1305, Austronesian, Oceanic); Niuean (niu; niue1239, Austronesian, Oceanic).

3.1.4. Dimension 4: order of subject (and negator) and verb, and NP orders

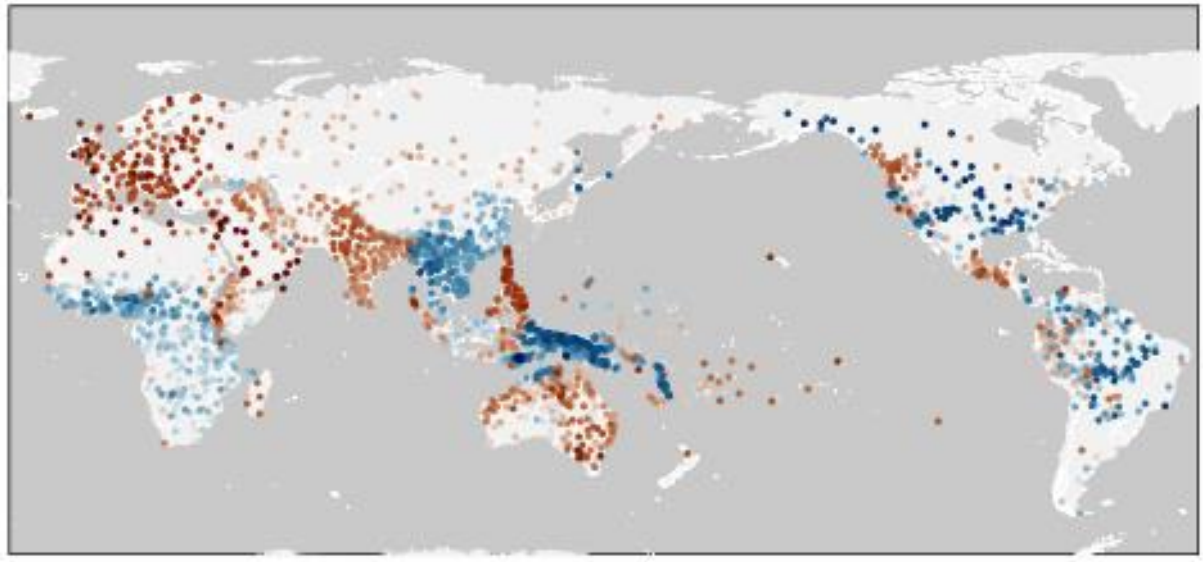
The other major aspect of clausal word order, the order of subjects and predicates, is found to have the strongest associations with both ends of Dimension 4. Clause-initial negation, which is overwhelmingly (but not exclusively) associated with verb-initial languages, also shows strong positive associations with Dimension 4. Unlike the word order correlations evident in Dimension 1, a number of NP-internal correlations are found with Dimension 4, leading to a number of languages which are not verb-initial nonetheless displaying high values on this dimension.

Direction	Feature	r^2
High	VSO order	0.22
	Clause-initial negation	0.20
	Numeral precedes noun	0.19
	Clause-initial Wh-question words	0.11
	Adjective precedes noun	0.11
	Genitive precedes noun	0.10
	Relative pronouns	0.10
	Clause-final negation	0.10
	Inalienable possession	0.10
	Object agreement prefix	0.14
	Numeral follows noun	0.14
Low	SV order	0.21

Table 7: Features characterising the extremities of Dimension 4.

The order of subject and verb again reflects Greenberg's classification of the world's languages by clausal word order. As with the order of object and verb, seen in Map 5, we can identify a number of contiguous areas which are high or low on this dimension. The relative paucity of VS languages, compared to SV languages, means that it is easiest to consider the distribution of these languages compared to a background of SV languages. The languages at the top of Dimension 4 are mostly Semitic and Berber languages from north Africa and the Middle East, and the Celtic languages of western Europe, though certain south-eastern Australian languages such as Warrnambool (gjm; warr1257, Pama-Nyungan, Kulinic), Wembawemba (xww; wemb1241, Pama-Nyungan, Kulinic) and Muk-Thang (Garnai) (unn; gana1278,

Pama-Nyungan, Gippsland) are also found at this extreme. Languages at the opposite extreme of this dimension are found in North America, including the Athabaskan languages Dena'ina (tfn; tana1289), Kaska (kkz; kask1239) and Slavey (xsl; sout2959), and the Siouan languages Lakhota (lkt; lako1247), Stoney (sto; 1ton1242), Hidatsa (hid; hida1246) and Hocąk (Winnebago) (win; hoch1243), as well as in languages from various families on the fringes of New Guinea, such as Puare (pux; puar1240) and Barupu (wra; wara1302) (Skou family); Riantana (ran; rian1263, (Trans New Guinea?), Kolopom), and the Timor-Alor-Pantar languages Tanglapui/Sawila (tpg; sawi1256), Lamma/Western Pantar (lev; lamm1241), Adang (adn; adan1251), Abui (abz; abui1241) and Kamang (woi; kama1365).



Map 8: Position of languages on Dimension 4 (blue = low, brown = high).

3.2. Geographically or genealogically recognisable regions

In this section we return to the dimension plots seen in Figure 3 (and compare also with Map 12), and examine recognisable geographic or genealogical regions to determine whether, and to what extent, they correspond to distinct 'regions' in typological space. To assess whether a given geographic or genealogical group clusters on one side of a given dimension, we perform a one-sided *t*-test comparing the values of languages within the group along that dimension, and the values of all languages in the dataset along that dimension. Generally, *t* values greater than 20 or less than –20 indicate that a group of languages shows extreme values along a given dimension, and/or forms a tight cluster along that dimension. A *p*-value close to zero

indicates that the means of the two populations being compared are significantly different; however, since *p* values are generally lower for larger datasets, the results should be interpreted on the basis of the *t* statistic as well as the *p* value. We also report the degrees of freedom (*df*) for each analysis.

3.2.1. Western Old World: Europe, Arabia and North Africa

As mentioned in 3.1.3, the languages of (western) Europe almost all occupy a position high on Dimension 3 (Western Old World) ($t = 40.24$, $df = 178.89$, $p < 0.001$, according to a one-sided *t*-test) and 4 ('order of subject (and negator) and verb, and NP orders'), and moderately low on Dimension 2 ('verbal elaboration') ($t = –14.39$, $df = 175.57$, $p < 0.001$). The region of typological space that can be seen in the combination of these two dimensions is quite separate from the rest of the cloud; exceptions to this separation, found much lower on Dimension 3, are recognised as outliers within Europe: Basque varieties (eus; basq1248), Hungarian (hun; hung1274), Gagauz Turkish (gag; gaga1249), and (to a lesser extent) the Celtic languages. The mixed word-order typology of most of the European languages (with head-initial parameters dominating at the clause level, and head-final parameters predominant within NPs) means they occupy a position in the middle of Dimension 1 ('order of object and verb'), and they can be seen to occupy a distinct, albeit interior, position in the plot of Dimension 1 vs. Dimension 2. In Figure 4 we can see that the languages of Europe occupy a compact region in typological space in each of the dimension plots, including those that do not involve dimensions 2 or 3, though they are not part of the 'fringe' of typological space.

Figure 5 shows the position of the languages of Arabia and North Africa; not as compact as the European languages, they can also be characterised as occupying a fringe positions on the plot of Dimensions 1 and 3 ($t = 20.49$, $df = 52.25$, $p < 0.001$; $t = 18.38$, $df = 49.89$, $p < 0.001$), and are higher on Dimension 1 than the European languages ($t = 18.42$, $df = 73.60$, $p < 0.001$), but not significantly higher on Dimension 4 (two-sided $t = 0.92$, $df = 53.17$, $p = 0.36$). The outliers at the lower end of Dimension 3 for this group of languages are mixed languages in the areas, such as Kumzari (zum; kumz1235, Indo-European (?), Iranian), between Arabia and Persia, and Kwarandzyey (/Korandje) (kcy; kora1291), a Songhai language spoken in the extreme north-east of the Sahara in a Berber linguistic environment.

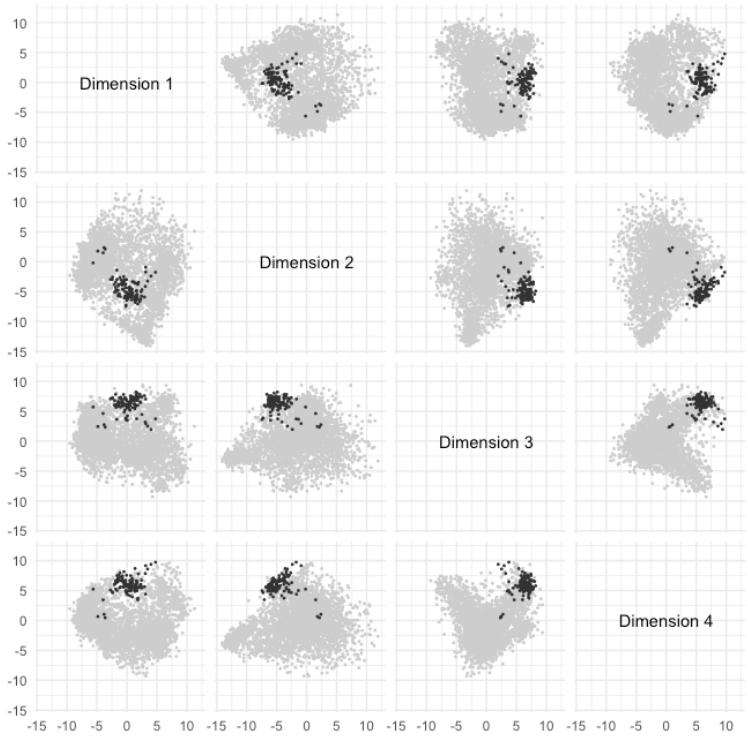


Figure 5: Languages of Europe highlighted on the dimension plots.

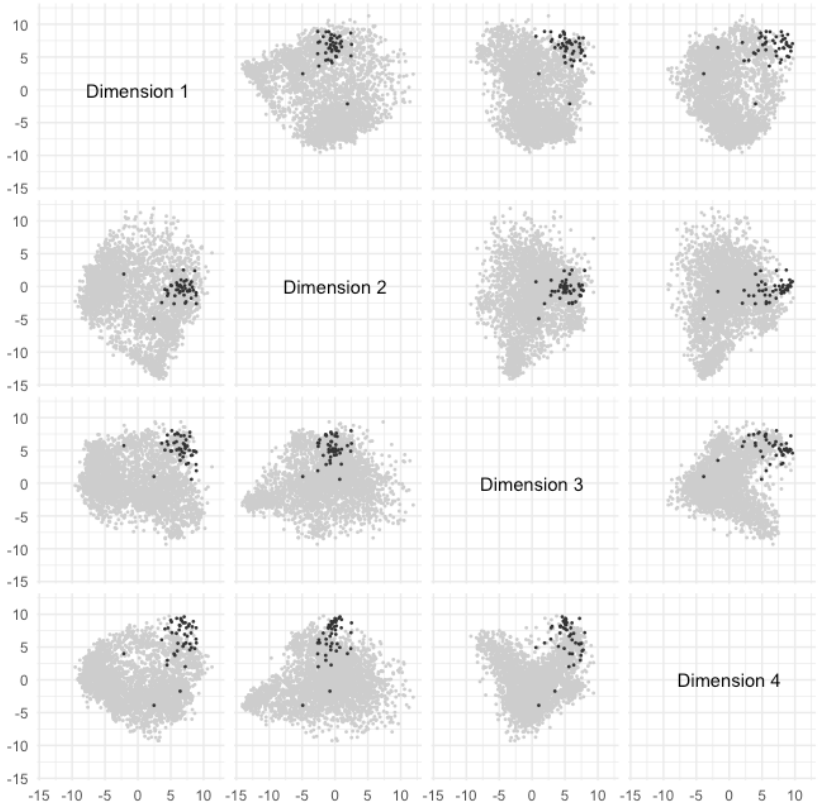


Figure 6: Languages of Arabia and North Africa highlighted on the dimension plots.

3.2.2. Mainland Southeast Asia

The languages of Southeast Asia represent a number of typologically convergent language families, all low on Dimensions 2 and 3 ($t = –41.17$, $df = 366.98$, $p < 0.001$; $t = –30.25$, $df = 821.84$, $p < 0.001$). The outliers for this group, typologically, are also outliers geographically. The most divergent languages are the Nungish languages of northern Myanmar and adjacent China, high on Dimension 2, and the Nicobarese languages of the Nicobar Islands, high on Dimension 4 (raising questions about their inclusion in a 'Mainland Southeast Asia' area). As with the languages of Europe, the languages of Southeast Asia largely cluster together even in plots that do not involve Dimension 2 or Dimension 3.

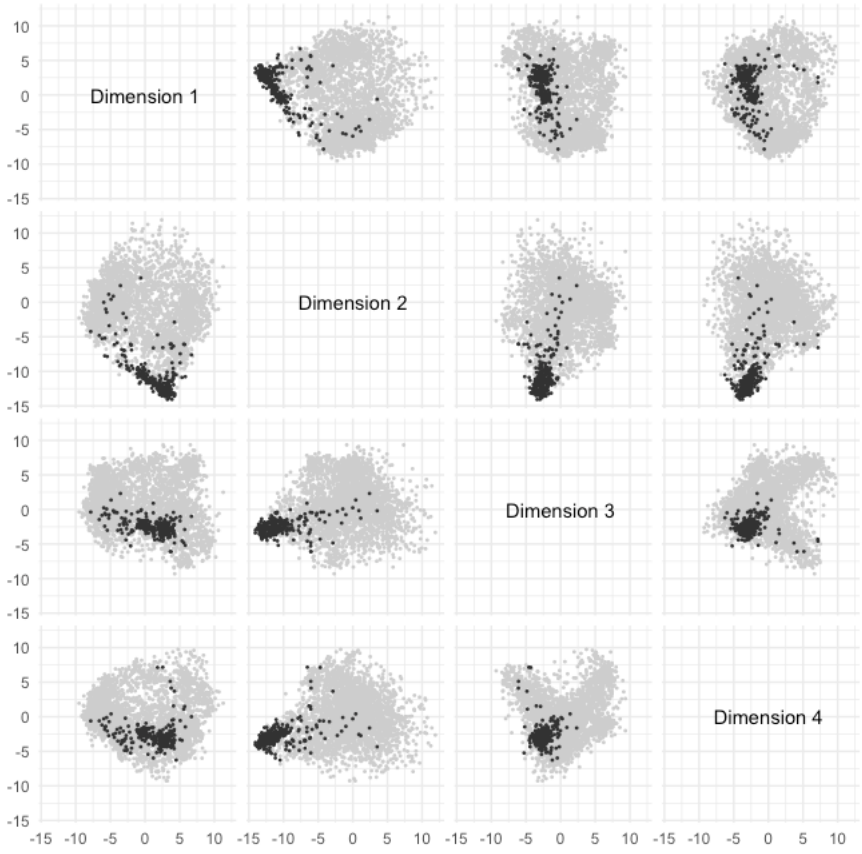


Figure 7: Languages of mainland Southeast Asia highlighted on the dimension plots.

3.2.3. Core South Asia

The Indic (< Indo-European) and Dravidian languages of South Asia also cluster together, though not at the periphery of any of plots, except for their low position on

Dimension 1 (t = –41.99, df = 205.58, p < 0.001), and relatively high position on Dimension 3 (t = 29.02, df = 136.18, p < 0.001). The typological outliers for this area, low on Dimension 3 or high on Dimension 1, are Vedda (ved; vedd1240, Indo-European, Indic), from Sri Lanka, and Dari (prs; dari1249, Indo-European, Iranian), the eastern variety of Farsi spoken in Afghanistan and not typologically assimilated to the languages of the region.

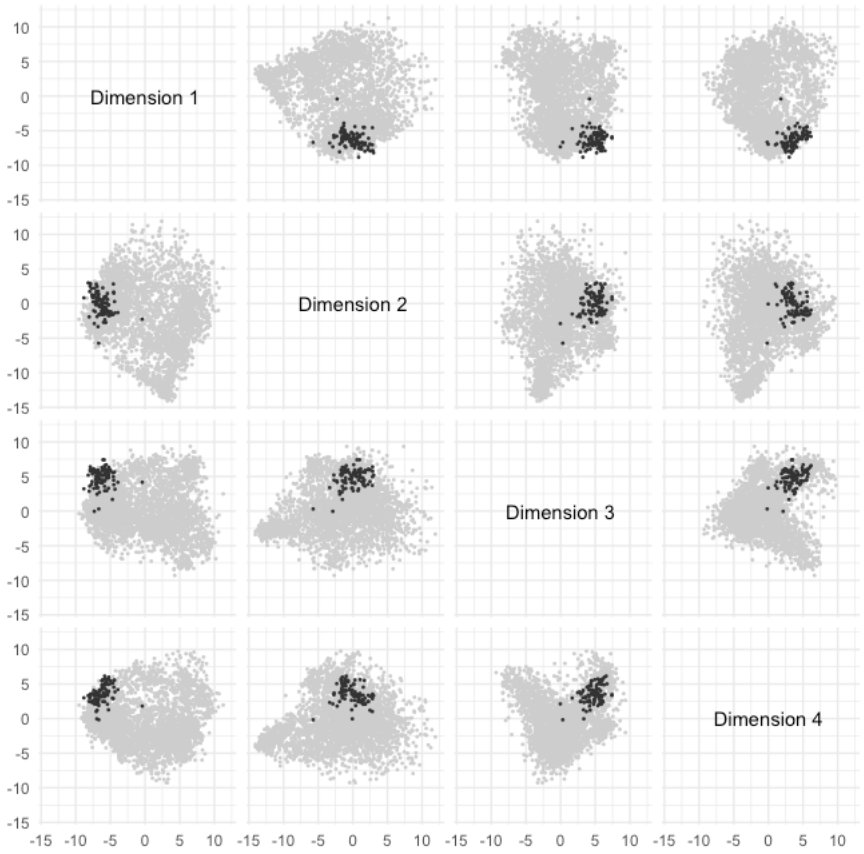


Figure 8: Languages of 'core South Asia' highlighted on the dimension plots.

3.2.4. Inner Asia

The core Eurasian profile of a radically head-final language (low on Dimension 1: t = –25.13, df = 191.91, p < 0.001) with a modest level of morphological elaboration (moderately greater than zero on Dimension 2: t = 4.52, df = 165.25, p < 0.001) is most strongly realised in Inner Asia, where Mongolic, Tungusic, Turkic and Uralic languages share many typological features. The outliers in this group are recently-

arrived varieties of Mandarin (Dungan, Urumqi and Taz)[25] and Arabic (Afghanistani Arabic and Bukhara Arabic)[26], which are low on Dimension 2 and high on Dimension 1, respectively.

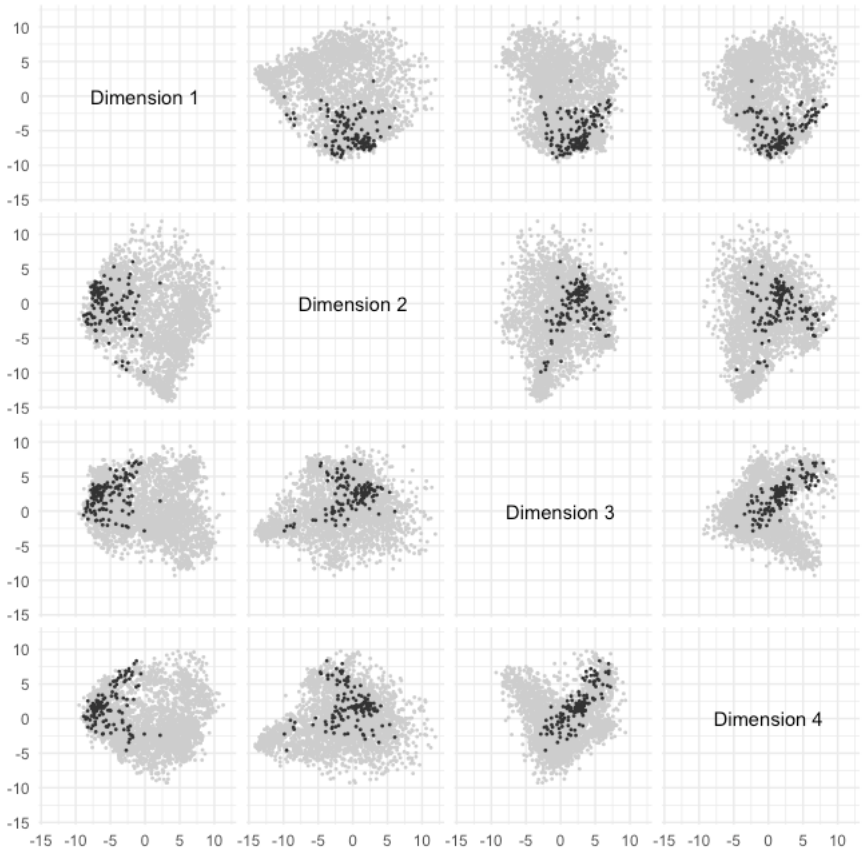


Figure 9: Languages of Inner Asia highlighted on the dimension plots.

3.2.5. North America

The languages of North America are widely dispersed, though the average position, and densest grouping, is both high on Dimension 2 (t = 24.8, df = 406.47, p < 0.001) and low on Dimension 3 (t = –17.3, df = 428.62, p < 0.001), indicating a head-marking, morphologically complex language that is maximally different from the languages of western Eurasia. In Dimension 1 and Dimension 4 there is no apparent pattern (two-sided t = –0.94, df = 322.14, p = 0.35), but in Dimension 1 the languages on average have values slightly greater than zero (t = 6.60, df = 339.84, p < 0.001).

[25] Dungan (dng; dung1253); Urumqi (cmn; wulu1243); Taz (cmn; north3283).

[26] Afghanistani Arabic (abh; taji1248); Bukhara Arabic (auz; uzbe1248).

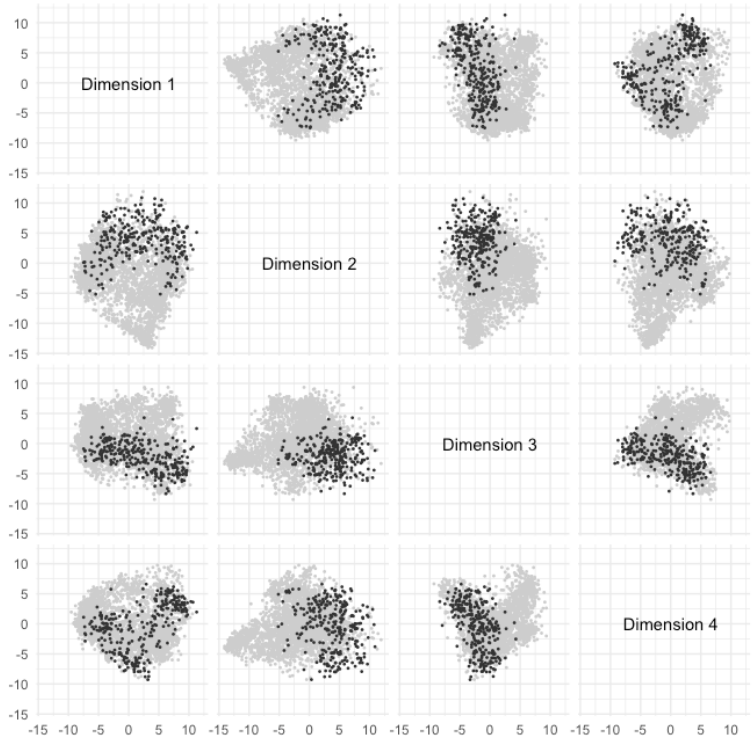


Figure 10: Languages of North America highlighted on the dimension plots.

3.2.6. Mesoamerica

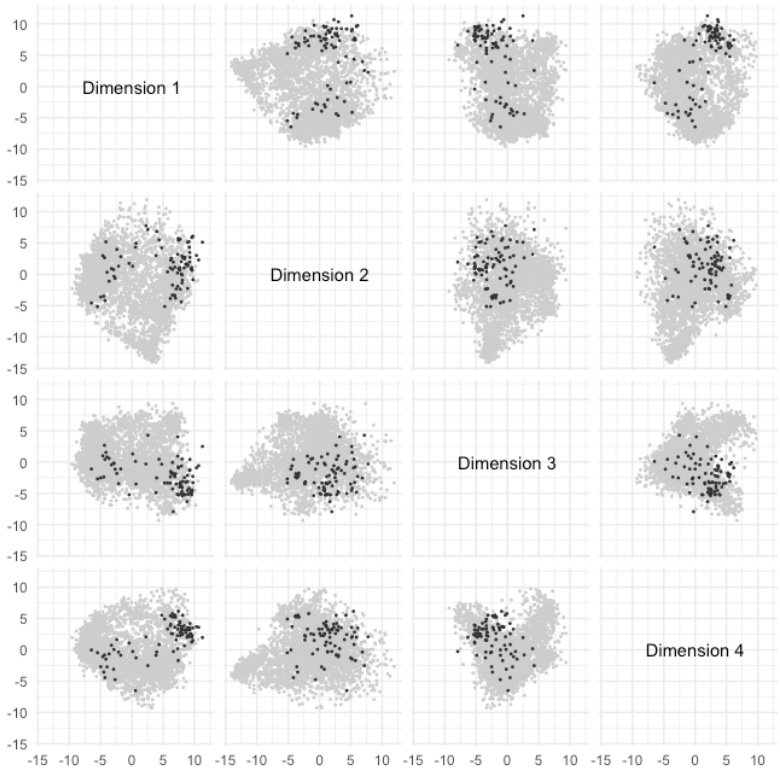


Figure 11: Languages of Mesoamerica highlighted on the dimension plots.

Focussing just on the languages of Mesoamerica as a sub-region within North America we find a high degree of typological dispersal, but with a cluster high on Dimension 1 ($t = 8.63$, $df = 86.28$, $p < 0.001$) and middling high on Dimension 4 ($t = 6.81$, $df = 89.38$, $p < 0.001$).

3.2.7. The Philippines and Taiwan

The 'Philippine-type languages' of the Philippines and Taiwan, which, while mostly Austronesian, do not form a valid subgroup within that family, can be found high on Dimension 4 ($t = 31.15$, $df = 56.38$, $p < 0.001$) and low on Dimension 3 ($t = –7.88$, $df = 64.48$, $p < 0.001$), where they form a fringe to typological space.

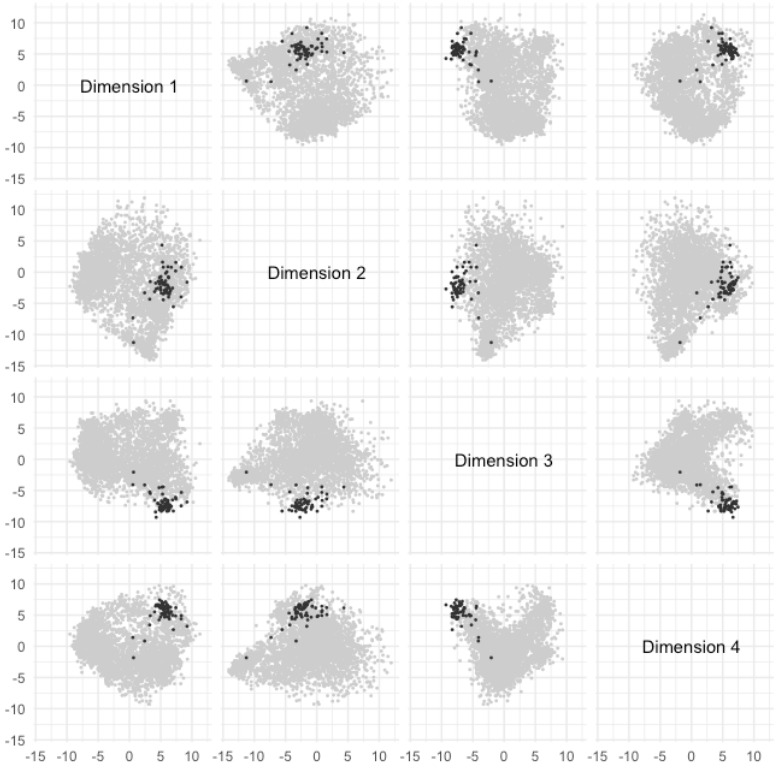


Figure 12: Languages of the Philippines and Taiwan highlighted on the dimension plots.

In other dimension plots they also form a tight cluster, with the divergent languages from this region (high or low on Dimension 2, or low on Dimension 1 or 4) being 1) the Austronesian languages of Taiwan (high on Dimension 2), 2) Iraya (iry; iray1237, Austronesian) from Mindoro in the Philippines, and Taiwanese (nan; taib1242, Tibeto-Burman, Sinitic), the intrusive Sinitic language of Taiwan (low on Dimension 2),

and 3) the southern Austronesian languages in this cluster, such as Talaud (tld; tala1285) and Sangir (sxn; sang1336) from northern Indonesia (low on Dimensions 1 and 4).

3.2.8. Eskimo-Aleut

The languages of the Eskimo-Aleut family are very low on both Dimension 1 ($t = –15.51$, $df = 15.79$, $p < 0.001$) and very high on Dimension 2 ($t = 39.45$, $df = 16.29$, $p < 0.001$), indicating a morphologically elaborate group of extremely SOV languages. As a small and young family they form a tight cluster, and represent an extreme extension of the North American (or North-east Asian) linguistic type.

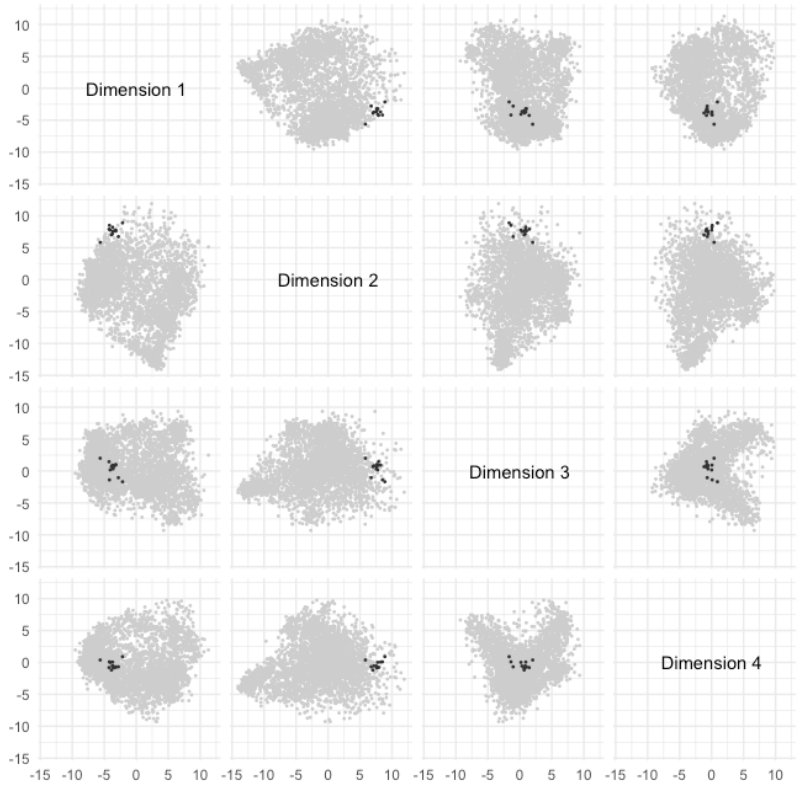


Figure 13: Languages of the Eskimo-Aleut family highlighted on the dimension plots.

3.2.9. North-west Caucasus

The languages of the North-west Caucasus family occupy a position similar to the Eskimo-Aleut languages, but more extreme (Dimension 1: $t = 5.24$, $df = 8.01$, $p < 0.001$; Dimension 2: $t = 5.47$, $df = 7.88$, $p < 0.001$).

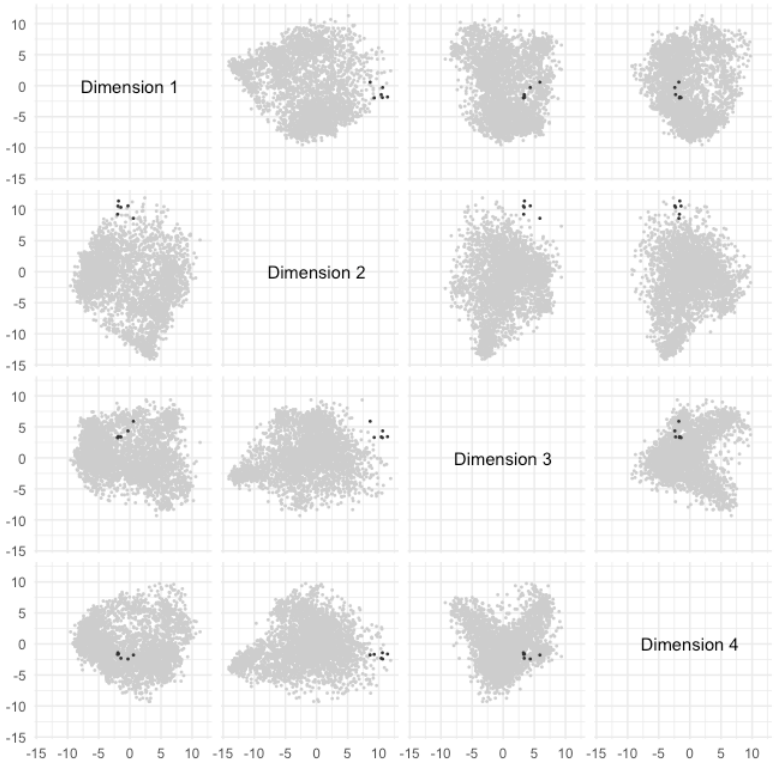


Figure 14: Languages of the Eskimo-Aleut family highlighted on the dimension plots.

3.2.10. Narrow Bantu

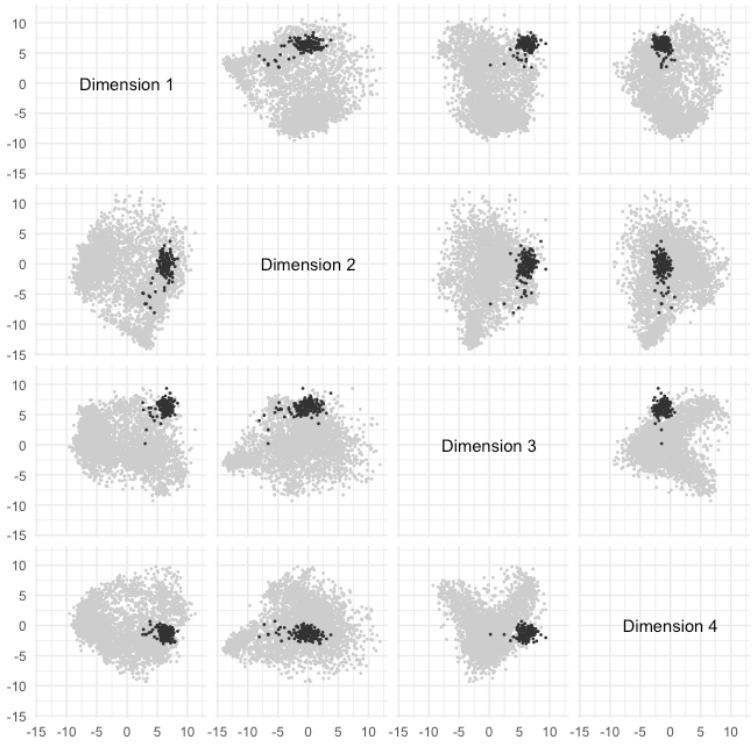


Figure 15: Languages of the Narrow Bantu subgroup highlighted on the dimension plots.

The many languages of the Narrow Bantu subgroup, found in a broad, contiguous range across southern Africa, are highly typologically congruent, being found high on Dimensions 1 and 3 ($t = 50.14$, $df = 587.47$, $p < 0.001$; $t = 51.67$, $df = 318.28$, $p < 0.001$), and low on Dimension 4 ($t = –12.09$, $df = 518.94$, $p < 0.001$). They represent a typological extension away from the rest of the language cloud, seen in the plot of Dimension 3 vs. Dimension 4. Typological outliers of this group (lower on Dimension 2, or lower on Dimension 3) include the peripheral Bantu languages from the north-west of the Bantu expanse, in Cameroon, The Congo, or the Democratic Republic of the Congo, which are more isolating than the 'modal' Bantu language.

3.2.11. Greater Abyssinia

The languages of Greater Abyssinia, centred around the Horn of Africa, are typologically diverse, but are all relatively higher on Dimensions 2 ($t = 5.25$, $df = 105.57$, $p < 0.001$), and lower on Dimension 3 ($t = –5.77$, $df = 84.27$, $p < 0.001$), than the Bantu languages.

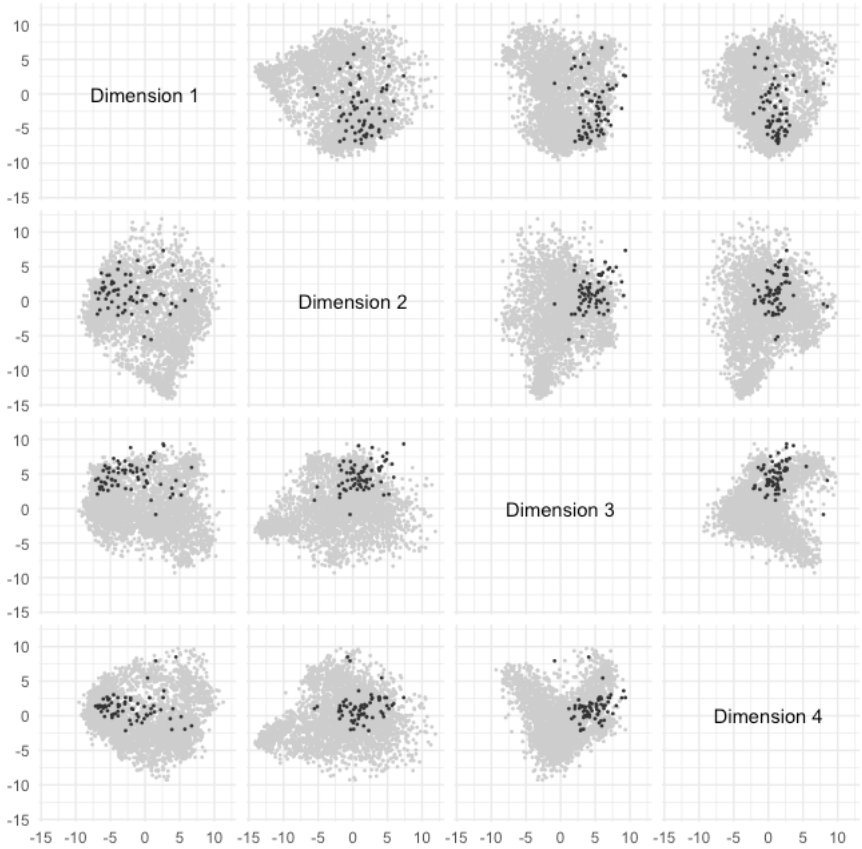


Figure 16: Languages of Greater Abyssinia highlighted on the dimension plots.

3.2.12. Pacific North-west

The languages of the Pacific North-west are typologically convergent languages from a number of families, high on Dimension 2 ($t = 17.36$, $df = 31.86$, $p < 0.001$), low on Dimension 3 ($t = –14.24$, $df = 30.02$, $p < 0.001$), and high on Dimension 4 ($t = 8.30$, $df = 29.04$, $p < 0.001$). The outliers lower on Dimensions 1 or 4 are at the northern or southern edges of the area (Tlingit (tli; tlin1245, Na-Dene, Tlingit) and Haida (hdn; nort2938, Haida), Klamath (kla; klam1254, Klamath-Modoc), Kalapuya (kyl; kala1400, Kalapuyan) and Molala (mbe; mola1238), respectively).

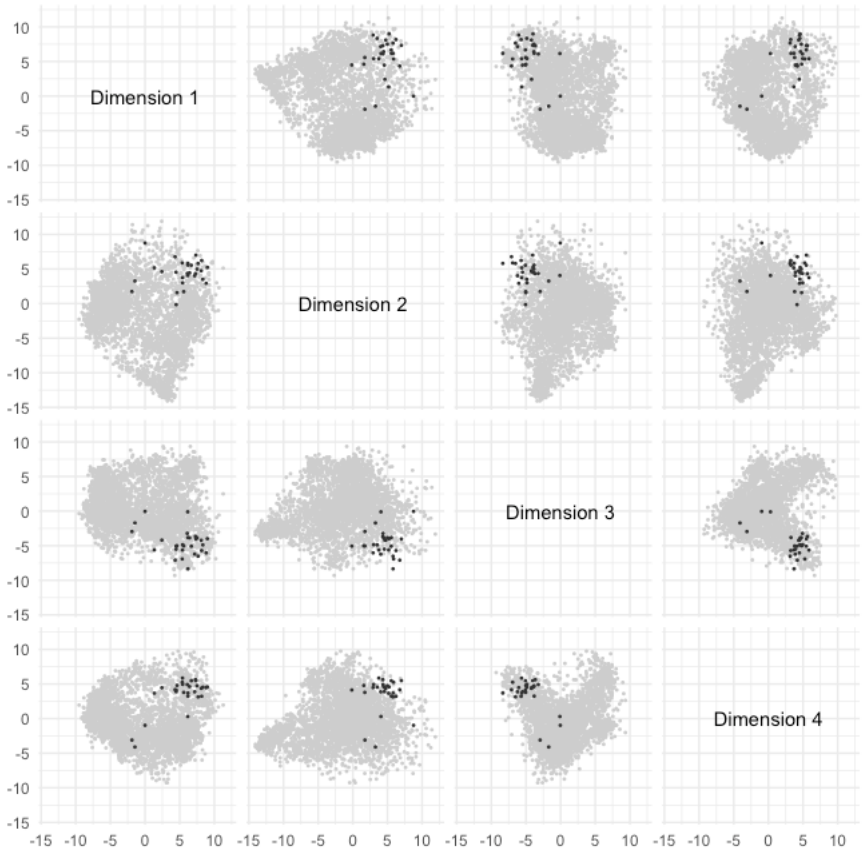


Figure 17: Languages of the Pacific North-west highlighted on the dimension plots.

3.2.13. Southern and Central Australia

The Pama-Nyungan languages of southern and central Australia occupy a small region of typological space which is low on Dimensions 1, 3, and 4 ($t = –27.94$, $df = 207.27$, $p < 0.001$; $t = –23.40$, $df = 183.71$, $p < 0.001$; $t = 15.38$, $df = 158.29$, $p < 0.001$), and in the middle of Dimension 2 (two-sided $t = –1.06$, $df = 182.96$, $p = 0.29$).

Languages higher on Dimension 2 are those which have some form of agreement, on the very or via clitics, and languages higher on Dimension 4 tend to be found in the south-east of the continent.

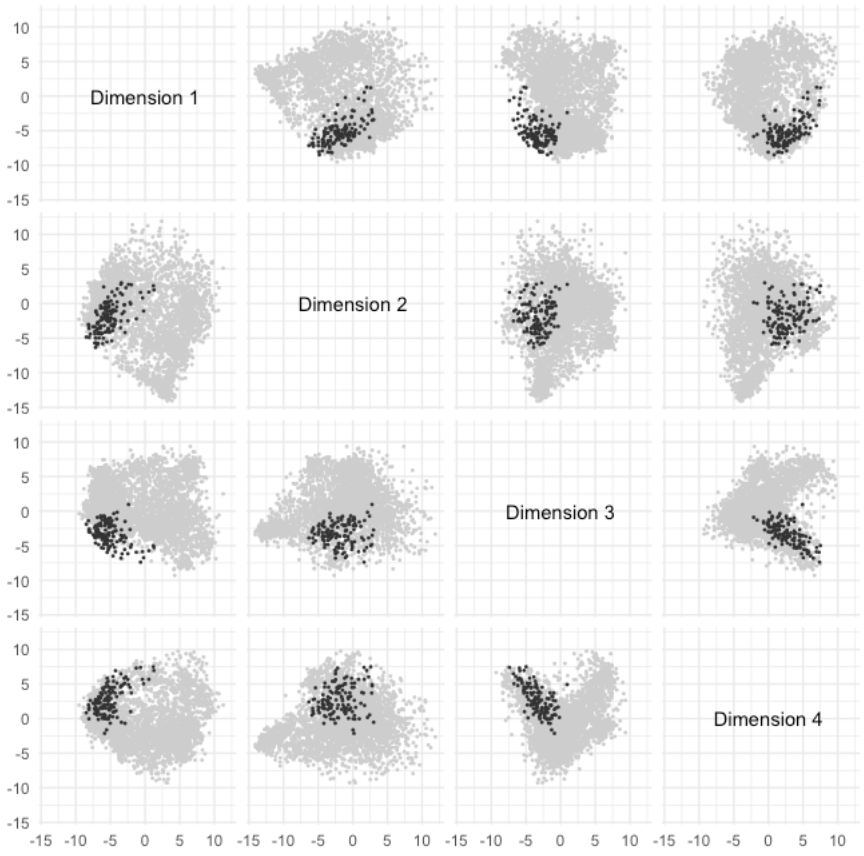


Figure 18: Languages of southern and central Australia highlighted on the dimension plots.

3.2.14. North-east Asia

The languages of North-east Asia, comprising the Ainu and Chukotko-Kamchatkan language families, as well as the Tungusic languages north of Hokkaido and the Eskimo-Aleut languages spoken west of the Bering Strait, and the Yukaghir languages. These languages are all high on Dimension 2 ($t = 11.59$, $df = 17.85$, $p < 0.001$), but do not occupy a typologically compact space in terms of the other three dimensions examined here (Dimension 1: two-sided $t = –2.28$, $df = 17.27$, $p = 0.035$; Dimension 3: two-sided $t = –1.08$, $df = 17.36$, $p = 0.29$; Dimension 4: two-sided $t = –0.54$, $df = 17.28$, $p = 0.59$).

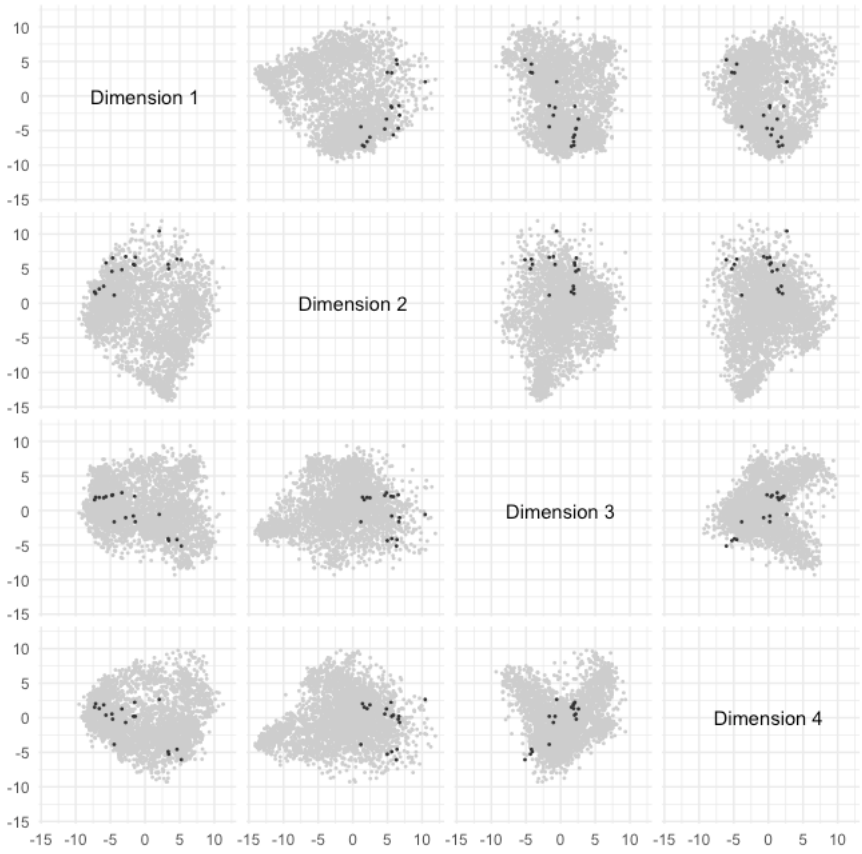


Figure 19: Languages of North-east Asia highlighted on the dimension plots.

3.2.15. Andes

The languages of the Andes are for the most part quite compact; exceptions are the isolates Camsá (kbh; cams1241), Esmeraldeño (atac1235), and Cholon (cht; chol1284), and to a lesser extent the Chibchan language Kuna (kvn; sanb1242), all in the north of the region except for Cholon, in Pre-Andine Peru. The main group of Andean languages, from the Aymaran, Barbacoan, Chocoan, Jivaroan and Quechuan families, are low on Dimension 1 ($t = –13.75$, $df = 45.60$, $p < 0.001$) and high on Dimension 2 ($t = 12.13$, $df = 44.46$, $p < 0.001$), and occupy middle positions on Dimensions 3 and 4 ($t = 2.80$, $df = 51.54$, $p < 0.01$; two-sided $t = 2.30$, $df = 44.66$, $p = 0.026$).

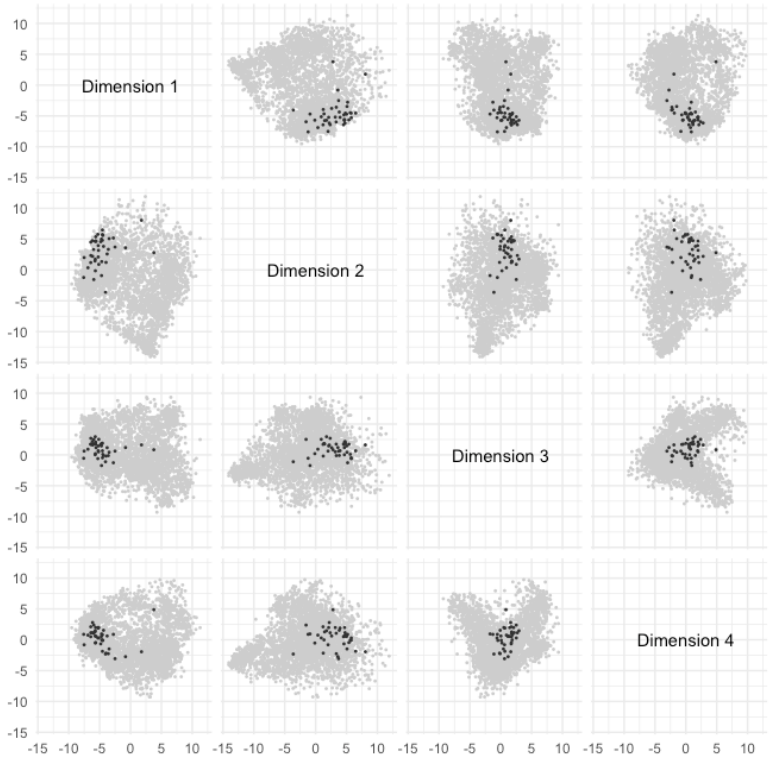


Figure 20: Languages of the Andes family highlighted on the dimension plots.

3.2.16. Mamoré–Guaporé

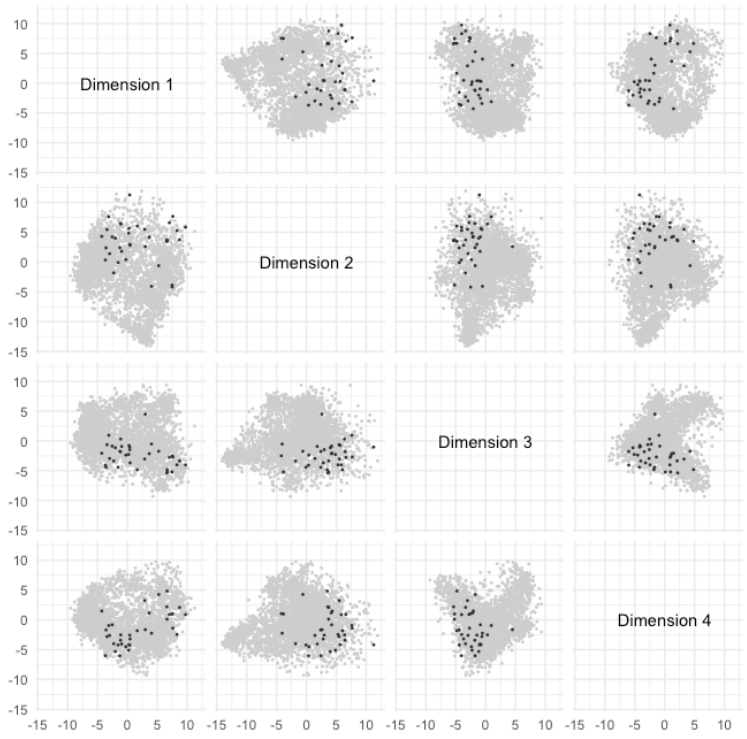


Figure 21: Languages of the Mamoré–Guaporé area family highlighted on the dimension plots.

The languages of the Mamoré–Guaporé area are typologically diverse in terms of the global dimensions of variation. In an analysis restricted to just South America these languages emerge as distinct, reflecting the widespread use of prefixal agreement and possessive affixes in these VO languages. As with other languages of the Americas, these languages are low on Dimension 3 ($t = –8.18$, $df = 37.66$, $p < 0.001$).

3.3. Universality? Macro-areas examined

In this section we examine whether the features that characterise the dimensions described in 3.1 are also relevant within individual macro-areas, paralleling the methodology advocated by Dryer (1989, and subsequent publications) that seeks confirmation for universals by their universal, independent attestation around the world. We have already seen (3.2.5) that the languages of North America occupy a position that does not cover the full extent of the dimensions, but in many dimensions does occupy the fringe, suggesting that the parameters of variation within this macro-area will be different, in at least some respects, from those that pertain to the globe as a whole.[27] In this section we report in outline the results of applying FAMD to individual macro-areas.[28] Tables 8 – 11 show the features that were discussed for each of the global dimensions of variation that were described in 3.1, with each row corresponding to a different macro-area, indicating, for each feature of the global FAMD, which dimensions (if any) of the local macro-area show associations with that feature (if any). For example, the difference between dominant OV vs VO order, a feature associated with Dimension 1 in the global analysis (3.1.1), appears in Dimension 1 in Africa, Eurasia, North America, the Pacific, and South America, but is relegated to Dimension 2 in Australia, where word order is less dominant a variable; in Africa VO order is also a feature with a strong association with Dimension 4. Case marking, also a feature of (global) Dimension 1 in 3.1.1, appears in Dimension 1 in all of the macro-areas except North America, where it is only found in the third dimension. In Eurasia case marking is associated with both the first and second

[27] In Appendix 6 similar plots are given to show the distribution of the languages of the other macro-areas in terms of the global variation.

[28] Note that the individual analyses result in different numbers of relevant dimensions, following the methodology described in section 2. For Africa eight dimensions emerged as relevant (though just the first two are sufficient to account for the variance in most of the languages); Eurasia, Australia and South America require four dimensions each, and for the Pacific and North American macro-areas three dimensions are optimal.

dimensions, indicating its importance in that continent. Gender, a correlate of Dimension 3 in the global analysis, appears in Dimension 1 in Africa, Eurasia and Australia, but is not correlated with any of the major dimensions in the Pacific or the Americas because of its rarity in those areas.

Importantly, most of the features we encountered in the global analysis are also relevant in (most of) the individual macro-areas, though their representation is increasingly scattered in the higher dimensions. As mentioned above, the order of object and verb is relevant in all macro-areas, though less prominently in Australia than in other areas; dependent-marking is relevant in five macro-areas, though less so in North America than in others. Agreement appears in the first two dimensions in all macro-areas, and the order of subject and verb is relevant to different degrees in all of the macro-areas; the scarcity of verb-initial languages in Africa and Eurasia lowers the relevance of this feature in the Old World. The different valency-adding devices (causatives and applicatives), which show a similarity to agreement in that they encode argument information on the verb, are generally less prominent in individual macro-areas, but are still relevant. The features associated with Dimension 3 globally are much less well-attested in individual macro-areas; this is to be expected, since, as we saw in 3.1.3, this dimension essentially presents as a cline across the Old World, and so is much less prominent from analyses of variation in the languages of Australia, the Pacific or the New World. Nonetheless, these languages from the western edge of the populated world are typological outliers, as seen earlier in Figure 3, in contrast to most of the languages of the Circum-Pacific region, and so this dimension must be part of a global investigation.

Features	**OV / VO**	**Case**	**Initial/final subordination**	**Prepositions**	**Agreement$_{prefix}$**	**(Agreement$_{suffix}$)**	**Obl V / V Obl**	**Postpositions**	**Total cases**
Africa	1,4	1		1	1,2	1,3	1	1	1
Eurasia	1	1,2	1	1	3		1		2
Pacific	1	1	1	1	2		1	1	
Australia	2	1	4		1	2			
North America	1	3	1	1	3	3	1	1	3
South America	1	1	1		1	2	1	1	

Table 8: Features associated with Dimension 1, and their positions in macro-areal analyses.

Features	Total agreement	Verbal synthesis	Modality affixes	Incorporation	Applicatives	Causatives	Possessive prefixes	Total tenses	SVO order	Symmetrical
Africa	1,2,7	2	2	4,5	6	2,4	7	8	1	2
Eurasia	2,3	2	1	4	3,4	2		2	2	2
Pacific	2,3	1	1		3		3	1	1	1,2
Australia	1,3,4	1		1,4	3	3		3		
North America	2,3			2	2	2	3			
South America	1,3				2	2	1	1	1	3,4

Table 9: Features associated with Dimension 2, and their positions in macro-areal analyses.

Features	Gender	Obligatory plural marking	$\text{Agreement}_{\text{suffix}}$	Relative pronouns	Ergativity	VOS order	Clusivity contrasts	Initial negation
Africa	1	1	1	4,5	6,8	3,4	3,6	3
Eurasia	1	1	1	1,3	1,2			1
Pacific			1			1	1	2
Australia	1	1	2		1,4	2		2
North America		2			1	1		1
South America		3,4	1				3	2

Table 10: Features associated with Dimension 3, and their positions in macro-areal analyses.

In summary, most of the groupings of features we identified in 3.1 can be justified in the context of the analysis of individual macro-areas, suggesting that these associations between features are likely to arise from universal properties of human language.

The following sections briefly discuss the features that appear in the FAMD analyses of individual macro-areas, including those which are not present in the global analysis.[29]

[29] A more detailed explication of the FAMD analysis of the individual macro-areas is shown in Appendix 5.

Features	Initial negation	VSO order	Gen/Adj/Num N	Relative pronouns	Final negation	Inalienability	N Num	SV order
Africa	3	3,6		4,5	2		7	1,3
Eurasia	1	1		1,3		3		1
Pacific	2	2				1		2
Australia	2	2				1		2
North America	1	1			3		1	1
South America	2	2			2	3,1	2	2

Table 11: Features associated with Dimension 4, and their positions in macro-areal analyses.

3.3.1. Africa

We can see in Table 12 that most of the features that determine variation within Africa are consistent with the parameters of global variation.[30] The differences that can be found involve the strong correlation of postpositions with SOV languages in Africa, which is not found globally, and the widespread use of prefixal plural marking, which is so common amongst the Bantu and other Niger-Congo languages that it plays a large role in the continent as a whole. In the fourth dimension, varieties of Malagasy (bhr; bara1369) are differentiated by the nature of its voice system (here dubbed 'superapplicative', following Naylor 1995), and in the fifth and sixth dimensions, which are justified following the same procedures described in Section 2, we find features that identify certain Chadic and South Semitic languages which display infixation, and a small number of mostly East Sudanic languages which have ergative patterns.[31]

[30] Note that, as discussed in Section 2, we consider the languages of northern Africa, north of the Sahara, to be part of the macro-area Eurasia, rather than Africa, for the reasons outlined there. As such Arabic and Berber languages are not included in the analysis of Africa separate from Eurasia.

[31] Dimensions 2 and 3 correspond very closely to Dimension 2 (3.1.2) and Dimension 4 (3.1.4) from the global analysis.

Dimension	Low			High
1	SVO	Agreement prefixes	Plural suffixes	SOV, Postpositions
2	Negative particle		Verbal agreement	Causatives
3	Subject-Predicate	Gender		Predicate-Subject
4	'Superapplicative'	Double causatives, VOS, Incorporation		Head marking
5	Noun-modifier orders		Incorporation	Infixes
6	Gender in 1/2 pronouns		Applicatives	Ergativity
7	Incorporation	Possessive suffixes	Double negation	Third agreement position
8	Possessive classes	Third agreement position	VSO order	Ergativity

Table 12: Relevant features: Africa.
'Low': features showing a negative correlation with the relevant dimension;
'High': features showing a positive correlation with the relevant dimension.

3.3.2. Australia

Australia is most at variance with global norms in terms of morphosyntactic variation. As seen above, word order is not a correlate of the first dimension of variation in Australia (though it is represented in the second dimension). The features correlating with the major dimension of variation in Australia correspond to the long-discussed Pama-Nyungan/non-Pama-Nyungan divide, with the north(-west)ern non-Pama-Nyungan languages displaying prefixal agreement on verbs, often with portmanteau subject/object morphemes, clusivity contrasts in bound morphology, and gender systems. Opposing this are the Pama-Nyungan languages that occupy most of the continent, which tend to be more dependent-marking, with ergative case marking, and typically lacking gender contrasts. The second dimension picks out languages, typically in the south-east of the continent, which are verb-initial and which employ pronominal bases to which a productive affix is added (Daniel 2013).

Dimension	Low			High
1	Ergativity	Suffixing	Gender, Clusivity	Prefixal agreement
2	SOV order		Pronominal bases	VOS order
3	Verb agreement		Causatives	Applicatives
4	Verb agreement	Subordinating suffixes	Incorporation	N Dem order

Table 13: Relevant features: Australia.

3.3.3. Eurasia

As with Africa, the mapping of Eurasia in Section 4 reveals a number of clearly separated areas. Unique features associated with dimension 1 include the contrast between isolating languages and tense-marking languages. The second dimension has a strong east-west distribution, with high values in the west, where word order is manipulated to form content questions, and relative pronouns are used as subordinators.

Dimension	Low			High
1	Tense, SOV	Case marking	VO	Isolating
2	Prenominal relative clause		Initial Wh-, subject suffixes, gender	Relative pronouns
3	Accusative pronouns		Applicatives	Prefixal agreement
4	VS			SV
5	Causatives			Ergativity

Table 14: Relevant features: Eurasia.

3.3.4. Pacific

In the Pacific we again see an OV vs. VO divide along the first dimension, correlating with suffixal subject morphology amongst the 'OV Papuan' languages and clusivity contrasts in the VO Austronesian languages. The second dimension introduces prefixal agreement as a major correlate.

Dimension	Low	High
1	SOV Subject agreement suffixes	SVO Clusivity contrasts
2	Prefixal agreement SV	Initial negation, VS Case marking
3		Applicatives, total agreement positions

Table 15: Relevant features: Pacific.

3.3.5. North America

Most of the features that appear in the analysis of North America are also present in the global analysis. Additionally, the second dimension distinguishes between prefixing and suffixing languages.

Dimension	Low	High
1	SOV	Negator-verb Verb-predicate
2	Suffixing Causatives	Prefixal agreement Prefixing
3	Applicatives Total agreement positions	

Table 16: Relevant features: North America.

3.3.6. South America

Dimension	Low	High
1	SOV Case marking	Prefixal agreement Prefixal possession
2	Symmetrical	Suffixal possession, object agreement Applicatives
3	Double negation Clusivity in bound morphology	Verb agreement Causatives
4	Modality affixes Tense	Suffixes plural marking Initial question particles
5	Applicatives	Dem N order

Table 17: Relevant features: South America.'Low': features showing negative associations with the relevant dimension; 'High': features showing a positive association with the relevant dimension.

South America shows the same VO vs. OV divide in the first dimension, but with less clear areality than is seen in other macro-areas. There is a strong Andean area defined by Dimension 2, abutting a Pre-Andine area defined by Dimension 3.

3.4. Correlations between dimensions?

By definition the different dimensions are as independent of each other as possible. Table 18 shows the overall correlations that can be found between the different dimensions; none of these correlations are significant, as shown in Table 18.

Dimension	1	2	3
1			
2	*0.034*		
3	*0.021*	0.025	
4	*0.001*	0.023	0.020

Table 18: Overall correlations between dimensions (r^2; negative correlations shown in italics).

Despite the different dimensions being overall independent, some correspondence is inevitable due to the presence of the same or similar features in more than one dimension.

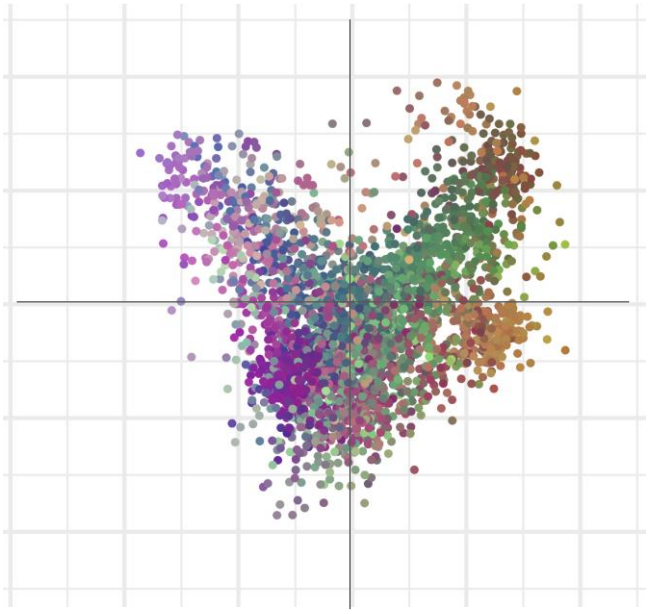


Figure 22: Dimension 3 (x) and Dimension 4 (y).

More interestingly, some dimensions correlate with others in just part of their range. Most dramatically, while there is only minimal correlation between Dimension 3 and Dimension 4 ($r^2 = 0.020$), the plot showing these two dimensions together (Figure 22, from Figure 3) shows clearly that at the lower half of Dimension 3 (on the left) there is a negative correlation with Dimension 4 ($r^2 = 0.290$), while at the upper half of Dimension 3 (on the right) there is a positive correlation with Dimension 4 ($r^2 = 0.258$). This reflects a split in the typology of verb-initial languages depending on whether they are found in the west (Celtic, Semitic) or the east (Philippines, Central America), since these languages are not typologically uniform. There is a weak positive correlation between Dimensions 1 and 4 in the upper half of Dimension 1 ($r^2 = 0.064$), reflecting the rarity of languages that are both OV and VS in their word order. We find a slightly higher positive correlation with Dimension 2 at the bottom tail (that is, for those values more than one standard deviation below the mean) of Dimension 1 ($r^2 = 0.094$); this corresponds to the fact that the very bottom of Dimension 1 is occupied by languages from the Himalayas (which are both low on Dimension 1, and concentrated in central Eurasia), which are less morphologically elaborate than many less 'extreme' SOV languages, and moving away from this edge almost inevitably leads to greater morphological elaboration, higher on Dimension 2. The lower half of Dimension 2 shows positive correlations with both Dimension 3 ($r^2 = 0.205$ in the extreme bottom) and Dimension 4 ($r^2 = 0.153$ in the lower half). These correlations largely reflect the position of the languages of Europe, high in Dimension 3 and Dimension 4, but in the lower half of Dimension 2. The last correlation we draw attention to is the lower half of Dimension 4, where we find a weak positive correlation with Dimension 3 ($r^2 = 0.085$) (see Figure 22).

4. Features with minimal contribution to global linguistic variation

In Section 3 we discussed the features that contribute to the dimensions that best describe global (and regional) morphosyntactic variation. This section briefly discusses some of the features that have the least contribution to global variation, either because they are so rare, they have a very limited distribution, or they appear in so many languages with little or no association with other parts of the language (at least, as far as is coded in the database used). Some of these features are listed in Table 19, which is not intended to be exhaustive.

Extremely Rare	Geographically Limited	Widely Ubiquitous
Polar questions formed by verbal reduplication	Genitive subjects	Predicative possession with a 'have' verb, or genitive subject
More than three agreement positions on the verb	Polar questions forms with word order change	Polar questions forms with particles or intonation
Marked absolutive case	Verb agreement by tone	Adnominal demonstrative identical to pronominal demonstratives
Incorporation of transitive subjects into verbs	Philippine-type voice systems	Presence of a perfective in the aspect system

Table 19: Different features with minimal contribution to global categories of variation.

Examples of some of the 'extremely rare' features are shown in (4) – (7); in Yao'an Lolo (ycl; lolo1259, Tibeto-Burman, Lolo-Burmese; Merrifield 2010) the only marker of the question is the reduplication of the verb (the only language in our database with this feature). In KinyaRwanda (Kimenyi 1980) we see a verb with five agreement positions filled on the verb; the database contains only 12 languages with more than three agreement positions. The Nias (nia; nias1242, Austronesian, Batak-Barrier Islands; Donohue and Brown 1999, Brown 2005, Donohue 2008) sentences show the alternation of the unmarked *ulö* 'snake' in an A function, and the marked *g-ulö* 'snake' in absolutive functions; sixteen other languages in the database have marked absolutive cases, most (11) of which also mark the ergative role. In Boni (/Aweer) (orm; awee1242, Afro-Asiatic, Cushitic, Omo-Tana; Sasse 1984) we see the rare case of an incorporated A (Sasse notes that while a natural translation involves the passive, the verbform in (7) is clearly transitive); only three other languages are known to us with this feature.

(4) Yao'an Lolo: reduplication on verbs marking polar questions

Ni	*pia*	*cir-cir*	*ho*	*ar?*
2SG	clothes	wash-RED	REAL	PFV

'Have you already washed the clothes?'

(5) Kinyarwanda: more than three agreement positions
Abáana ba-zaa-ha-ki-mu-b-eerek-er-a.
children they-FUT-there-it-him-them-show-BEN-ASP
'The children will show it to him for them there.'

(6) Nias: marked absolutive case

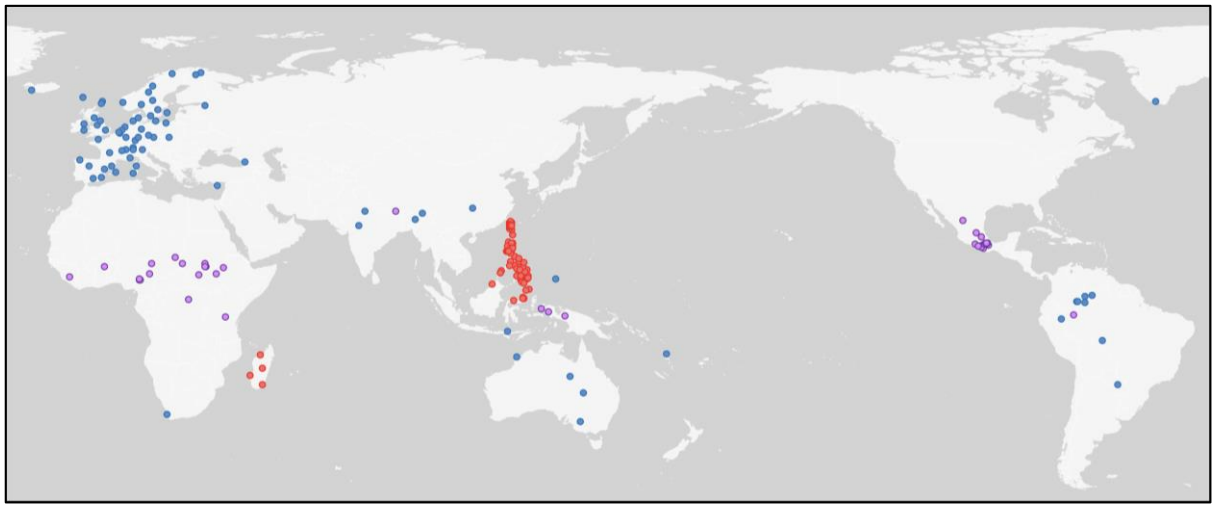
a. *I-usu* *g-ulö* *asu hö'ö.*
3SG.ERG-bite ABS-snake dog DIST
'That dog bit the snake.'
b. *I-usu* *n-asu ulö* *hö'ö.*
3SG.ERG-bite ABS-dog snake DIST
'That snake bit the dog.'
c. *Möi* *ga* *g-ulö.*
go here ABS-snake
'The snake came this way.'

(7) Boni: incorporated transitive subject
Míŋ *qwęęrą* *kawáyd'aadéed'i* *idohóoˆd'isa.*
house Boni/GEN usually womenˆbuild/IMPERF/3SG.M
'Boni houses are usually built by women.'

Map 9: Languages with extremely rare features.
Blue = more than three agreement positions; Red = marked absolutive case; Purple = both the preceding features; Black = polar questions marked by reduplication.

The distributions of languages displaying the first three of these features are shown in Map 9, where red indicates languages with marked absolutives, blue indicates languages with more than three agreement positions on the verb (compare with Map 1, which shows the total sample examined).

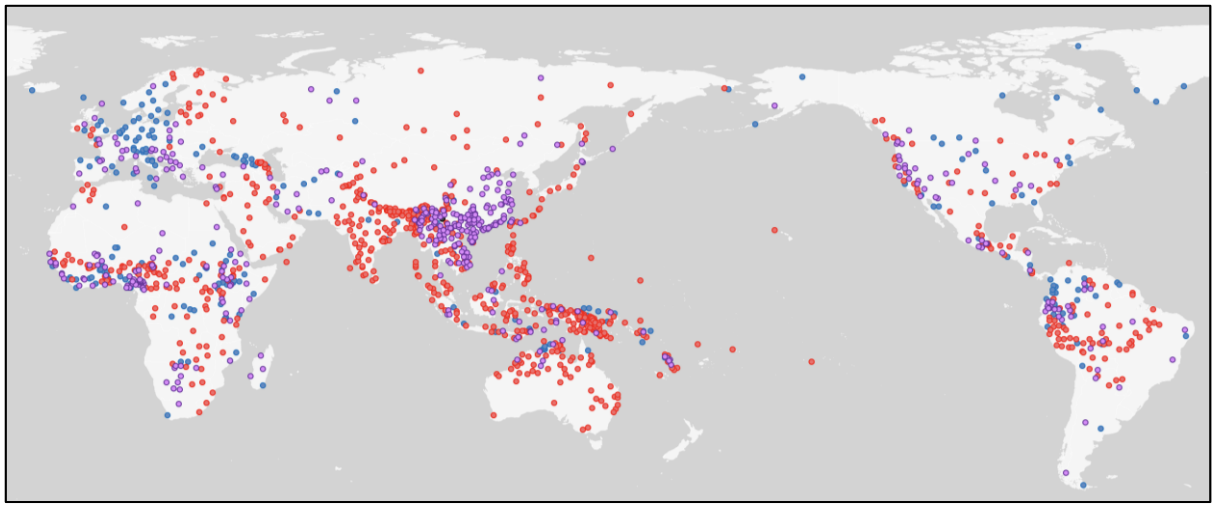
Examples of features that are more common than those shown in Map 9, but which have strong geographic concentrations, are shown in Map 10. While a number of parts of the world have languages in which word order changes in polar questions, the concentration in western and northern Europe is striking. Languages with Philippine-type voice systems are largely restricted to the Philippines and Taiwan, with the outlier group in Madagascar reflecting the migration from Southeast Asia ca 1,500 years ago. Languages which have tone as an exponent of verbal agreement are concentrated in Central America and in Central Africa.



Map 10: Features with geographically restricted ranges.
Blue = word order changes in polar questions;
Red = languages with Philippine-type voice systems;
Purple = tone as an exponent of verbal agreement.

Two features listed as widely ubiquitous in Table 19, the existence of a 'have' verb in the language, or the use of a particle to mark polar questions, are shown in Map 11. Both of these features are found across the map, though with different frequencies in

different continents. They are so widespread that they have very little probative value in understanding global morphosyntactic variation.[32]



Map 11: Features with widely ubiquitous distributions.
Blue = language includes a 'have' verb;
Red = language uses a particle to mark polar questions;
Purple = language has both a 'have' verb, and a polar question particle.

5. Conclusions

Without explicitly setting out to do so, our study has quantitatively confirmed many of the insights of 20th-century typological research concerning the main dimensions of morphosyntactic variation, by finding them as emergent properties of a bottom-up investigation of a large body of morphosyntactic data. We have shown that much of the variation between languages, both globally and within macro-areas, can indeed be largely explained by established typological parameters, as described in 3.1: the order of subjects and objects with verbs, dependent-marking settings, and the position of genitives, numerals and adjectives with respect to the nouns that they modify. The features that correlate with Dimensions 2 and 3, head-marking settings, verbal elaboration, and a number of features that are reminiscent of 'Standard Average European', have not all been proposed as factors underlying typological variation, but have been demonstrated here to be as important as more familiar word order

[32] A glance at Map 11 raises the suspicion that these might be relevant features at local levels; the distribution of 'have' verbs in South America, for instance, appears to be concentrated in the north-west.

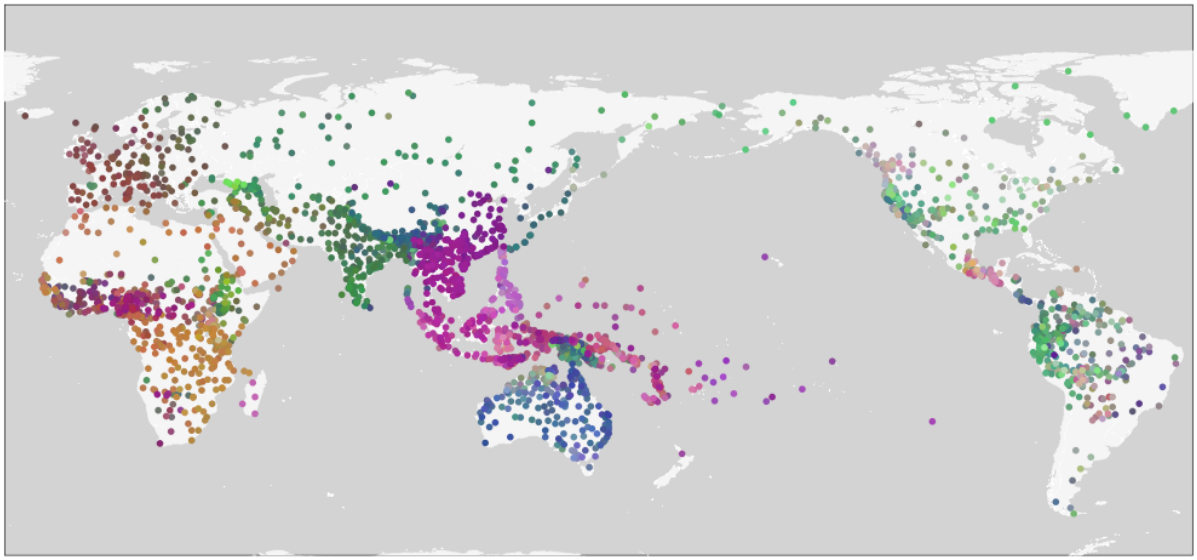
universals. The kinds of features described in 3.1 are summarised in Table 20; we can see that the main correlates of these dimensions are evenly split into those involving word order, and those involving morphology, with different dimensions addressing different kinds of word order or morphology. Secondarily we find that the presence of case-marking morphology, or the lack of extensive morphological processes in the language, or else word order parameters more relevant to Noun Phrase-internal elements, are also significant factors in assessing global morphosyntactic variation.

Dimension	Main type of feature	Secondary types of features
1	Word order (clausal; object)	Dependent marking
2	Verbal morphology, head marking	Isolating profile
3	Nominal morphology, gender	Ergativity, Clusivity
4	Word order (clausal; subject)	Word order (Noun Phrase)

Table 20: Different features with minimal contribution to global categories of variation.

In Map 12 we see the 3089 languages of the database, coloured according to the position of each language on Dimensions 1–3, with these dimensions mapped to red, green and blue colour components, as described in 3.1. Each language is then represented with a dot of the same colour that was seen in Figure 3, so that in Map 12 the colouring solely represents the position of the individual languages in typological space as defined by the first three dimensions – that is, the colour scheme is an emergent property of the linguistic analysis, and in no way involved any phylogenetic information, and only involved reference to geography in that AUTOTYP areas were used as controls in the analysis.[33]

[33] This means that Dimension 4 is not represented in the colouring in Map 12. Dimension 4 is shown in Map 8, and in Appendix 8 alternatives to Map 12, and Figure 3, are shown with colourings representing different combinations of dimensions, including Dimension 4.



Map 12: Emergent areality around the world. Map coloured according to the first three dimensions of variation, as discussed in 3.1, and shown in Figure 3

A number of regions are clearly identified by both geography and typology, as discussed in 3.2. Europe (e.g., Haspelmath 2001) is represented with a distinct brown colour, which is largely due to the languages being moderately low on Dimension 2, and high on Dimension 3 (compare Figure 3 in Section 3.1 and Figure 4 in Section 3.2.1). The isolating languages of Southeast Asia (e.g., Enfield 2005) are shown in magenta, as they are relatively high on Dimension 1, low on dimension 2, and moderately low on Dimension 3 (compare Figure 3 and Figure 6 in 3.2.2). Similar, though not as extreme, colours are found in the Macro-Sudan Belt (Güldemann 2009) in west-central Africa, and in Polynesia (e.g., Krupa 1982). Most of Australia (e.g., Bowern 2006, Dixon 2017) is in dark blue, as the languages there are low on Dimension 1 and moderately low on Dimension 2. The only bright green regions, high on Dimension 2, are found in the Caucasus (Catford 1977) and parts of North America (e.g., Mithun 1999). In North America the Pacific North-west (Mithun 2010) stands out, coloured in grey and having much higher values on Dimension 1; a similar typology is evident in the Oaxaca area in Central America, and the Mamore-Guapore region of South America (Campbell et al. 1986; Crevels & van der Voort 2008; see 3.2.16). The orange colour found in much of Sub-Saharan Africa represents the Niger-Congo Bantu languages, high in Dimensions 1, 2 and 3 (and low in Dimension 4, though this is not apparent from the colouring on the map); see 3.2.10. Separated by the Macro-Sudan belt, the verb-initial languages of North Africa and Arabia appear in dark orange, reflecting their position at the top of Dimension 4, but also high on

Dimension 3 (3.2.1). The languages of Taiwan and the Philippines are also high in Dimension 4, coloured in mauve following their position low on Dimension 3 (3.2.7).[34] The densely occupied space low in Dimension 1 and middling in Dimensions 2 and 3 is coloured in dark green, and spans the Eurasian steppe and South Asia (Janhunen 2023, Emeneau 1956, and many others; see 3.2.3 and 3.2.4); higher on Dimension 2, but otherwise in a similar position, the languages of the Ethiopian linguistic areas (Crass 2009) are coloured in a slightly lighter shade of green, and a darker green is found for the languages of Japan and Korea, representing a position lower on Dimension 2. Many of the languages of the Andes in South America are reminiscent of this pan-Eurasian typology (Constenla Umaña 1991, Adelaar 2009, Michael et al. 2012; see 3.2.15). In addition to these previously-discussed regions in typological space, we can also identify a number of emergent areas on the map, such as South-west China, and North-west Australia, Oaxaca (within Meso-America; 3.2.6), the Kimberleys in Australia, and the South-east Amazon, all clearly identifiable on Map 12.[35]

We mentioned in Section 2 our decision to use nested geographic areas, rather than genealogies, as controls. While most 'controls' in recent linguistics studies are based on genealogies, we have based our work on culturally-defined areas, specifically a set of 25 areas slightly modified from the AUTOTYP areas, as described in Section 2.

Our results not only confirm typologists' intuitions about the features that are most important for typological classification, but also show the efficacy of a bottom-up approach to the detection and mapping of areal patterns in morphosyntax.[36] The success (in terms of interpretable results) of the use of a large set of linguistic features, without any cherry-picking, shows that a holistic (or even 'super-holistic') approach to language typology (following, e.g., Ramat 1986, Plank 1998, Comrie 1988, 2001, and others) is a valid way to objectively assess claims about linguistic areality or linguistic universals.

[34] Taken with the grey areas discussed in the Americas, and the mauve from the Philippines, the languages of North Africa-Arabia represent a third verb-initial linguistic 'types', with a large number of typological features not associated with the verb-initial parameter.

[35] A higher-resolution version of this map can be found in Appendix 7.

[36] It has been suggested that we attempt a similar analysis using the Grambank database. This research has already been performed (Skirgård et al. 2023), and, owing to the different and smaller set of languages coded for a different and smaller set of features, the results are very different, though we note that Skirgård et al. 2023 also appear to have identified word order, head/dependent marking and gender as relevant to their analysis.

Acknowledgments

We thank two anonymous referees, whose input has greatly improved this paper, and the input from the editorial team that similarly contributed to the clarity of our presentation. We are grateful for their feedback.

Abbreviations

1 = 1st person	BEN = benefactive	GEN = genitive
2 = 2nd person	Dem = demonstrative	IMPERF = imperfective
3 = 3rd person	DIST = distal	M = masculine
ABS = absolutive	ERG = ergative	PFV = perfective
ASP = aspect	FUT = future	REAL = realis

References

Adelaar, Willem F. H. 2009. *The Languages of the Andes*. Cambridge: Cambridge University Press.

van der Auwera, Johan. 2011. Standard Average European. In Bernd Kortmann & Johan van der Auwera (eds.), *The Languages and Linguistics of Europe: A Comprehensive Guide*, 291–306. Berlin: de Gruyter Mouton.

Bickel, Balthasar. 2002. The AUTOTYP research program. Invited talk given at the Annual Meeting of the Linguistic Typology Resource Center Utrecht, September 26–28, 2002.

Bickel, Balthasar & Johanna Nichols. 2006. Oceania, the Pacific Rim, and the Theory of Linguistic Areas. *Proceedings of the 32nd annual meeting of the Berkeley Linguistics Society*, 3–15. Berkeley Linguistics Society and the Linguistic Society of America.

Bickel, Balthasar & Johanna Nichols. 2013. Inflectional Synthesis of the Verb. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*, Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/22.)

Bickel, Balthasar, Johanna Nichols, Taras Zakharko, Alena Witzlack-Makarevich, Kristine Hildebrandt, Michael Rießler, Lennart Bierkandt, Fernando Zúñiga & John B. Lowe. 2023. The AUTOTYP database (v1.1.1). https://doi.org/10.5281/zenodo.7976754.

Bowern, Claire. 2006. Another Look at Australia as a Linguistic Area. In Yaron Matras, April McMahon & Nigel Vincent (eds.), *Linguistic Areas*, 244–265. London: Palgrave Macmillan.

Brown, Lea. 2005. Nias. In Alexander Adelaar & Nikolaus Himmelmann (eds.), *The Austronesian Languages of Asia and Madagascar*, 562–589. London: Routledge.

Bugaeva, Anna, Johanna Nichols & Balthasar Bickel. 2021. Appositive possession in Ainu and around the Pacific. *Linguistic Typology* 26(1). 43–88.

Campbell, Lyle. 1997. *American Indian languages: The historical linguistics of Native America*. New York: Oxford University Press.

Campbell, Lyle, Terrence Kaufman & Thomas Smith-Stark. 1986. Meso-America as a linguistic area. *Language* 62(3). 530–558.

Catford, John C. 1977. Mountain of tongues: the languages of the Caucasus. *Annual Review of Anthropology* 6. 283–314.

Cattell, Raymond B. 1966. The Scree Test for The Number of Factors. *Multivariate Behavioral Research* 1(2). 245–276.

Clauson, Gerard. 1956. The case against the Altaic theory. *Central Asiatic Journal* 2. 181–187.

Comrie, Bernard. 1988. Linguistic Typology. *Annual Review of Anthropology* 17. 145–159.

Comrie, Bernard. 2001. Different views of language typology. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language typology and language universals: An international handbook, Vol. 1*, 24–39. Berlin: Walter de Gruyter.

Constenla Umaña, Adolfo. 1991. *Las lenguas del área intermedia: introducción a su estudio areal*. San José: Editorial de la Universidad de Costa Rica.

Crass, Joachim. 2009. Ethiopia. In Bernd Heine & Derek Nurse (eds.), *A Linguistic Geography of Africa*, 228–250. Cambridge: Cambridge University Press.

Crevels, Mily & Hein van der Voort. 2008. The Guaporé-Mamoré region as a linguistic area. In Pieter Muysken (ed.), *From Linguistic Areas to Areal Linguistics*, 151–179. Studies in Language Companion Series. Vol. 90. Amsterdam: John Benjamins.

Croft, William. 2003. *Typology and Universals*. Cambridge: Cambridge University Press.

Daniel, Michael. 2013. Plurality in Independent Personal Pronouns. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*.

Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/35.)

DeLancey, Scott & Victor Golla. 1997. The Penutian hypothesis: Retrospect and prospect. *International Journal of American Linguistics* 63(1). 171–202

Dimmendaal, Gerrit J. 2011. *Historical Linguistics and the Comparative Study of African Languages*. Amsterdam: John Benjamins.

Dixon, R. M. W. 2017. The Australian Linguistic Area. In Alexandra Y. Aikhenvald & R. M. W. Dixon (eds.), *The Cambridge Handbook of Linguistic Typology*, 624–650. Cambridge: Cambridge University Press.

Donohue, Mark. 2008. Semantic alignment systems: what's what, and what's not. In Mark Donohue & Søren Wichmann (eds.), *Semantic alignment: typological and descriptive studies*, 24–75. Oxford: Oxford University Press.

Donohue, Mark & Lea Brown. 1999. Ergativity: some additions from Indonesia. *Australian Journal of Linguistics* 19(1). 57–76.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Language* 68(1). 81–138.

Dryer, Matthew S. 2013. Against the six-way order typology, again. *Studies in Language* 37. 267–301.

Dryer, Matthew S. & Martin Haspelmath (eds.) 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info, Accessed on 2023-01-06.)

Emeneau, Murray. 1956. India as a Linguistic Area. *Language* 32 (1). 3–16.

Enfield, Nicholas J. 2005. Areal Linguistics and Mainland Southeast Asia. *Annual Review of Anthropology* 34(1). 181–206.

Gabelenz, Georg von der. 1901. *Die Sprachwissenschaft: Ihre Aufgaben, Methoden und bisherigen Ergebnisse*. Leipzig: Chr. Herm. Tauchnitz.

Greenberg, Joseph. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph Greenberg (ed.), *Universals of Language*, 73–113. London: MIT Press.

Güldemann, Tom. 2009. The Macro-Sudan belt: towards identifying a linguistic area in northern sub-Saharan Africa. In Bernd Heine & Derek Nurse (eds.), *A Linguistic Geography of Africa*, 151–185. Cambridge: Cambridge University Press.

Guzmán-Naranjo, Matías & Laura Becker. 2021. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670.

Hammarström, Harald & Mark Donohue. 2014. Some principles on Macro-Areas in Typological Comparison. *Language Dynamics and Change* 4(1). 167–187.

Haspelmath, Martin. 2001. The European linguistic area: Standard Average European. In Martin Haspelmath, Ekkehard König, Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals, Vol. 2*, 1492–1510. Berlin: De Gruyter.

Humboldt, Wilhelm von. 1836. *Uber die Verschiedenheit des menschlichen Sprachbaues und ihren Einfluß auf die geistige Entwickelung des Menschengeschlechts*. Berlin: Druckerei der Koniglichen Akademie der Wissenschaften.

Janhunen, Juha A. 2023. The Unity and Diversity of Altaic. *Annual Review of Linguistics* 9(1). 135–154.

Josse, Julie & François Husson. 2016. missMDA: A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software* 70(1). 1–31.

Kari, James & Ben Potter (eds.). 2010. *The Dene-Yeniseian connection*. Anthropological Papers of the University of Alaska, New Series 5 (1-2). Fairbanks: Alaska Native Language Center.

Kimenyi, Alexandre. 1980. *Relational Grammar of Kinyarwanda*. University of California Publications, Linguistics, 91. Berkeley: University of California Press.

Krupa, Viktor. 1982. *Polynesian Languages: a survey of research*. London: Routledge and Kegan Paul.

Lê, Sébastien, Julie Josse & François Husson. 2008. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* 25(1). 1–18.

Macklin-Cordes, Jayden & Erich Round. 2022. Challenges of sampling and how phylogenetic comparative methods help, with a case study of the Pama-Nyungan laminal contrast. *Linguistic Typology* 26(3). 533–572.

de Menocal, Peter, Joseph Ortiz, Tom Guilderson, Jess Adkins, Michael Sarnthein, Linda Baker & Martha Yarusinsky. 2000. Abrupt onset and termination of the African Humid Period. *Quaternary Science Reviews* 19(1–5). 347–361.

Merrifield, Judith Thomas. 2010. *Yao'an Lolo Grammar Sketch*. Dallas: Graduate Institute of Applied Linguistics. (MA thesis.)

Michael, Lev, Will Chang & Tammy Stark. 2012. Exploring phonological areality in the circum-Andean region using a Naive Bayes Classifier. *Language Dynamics and Change* 4(1). 27–86.

Mithun, Marianne. 1999. *The Languages of Native North America*. Cambridge: Cambridge University Press.

Mithun, Marianne. 2010. Contact and North American Languages. In Raymond Hickey (ed.), *The Handbook of Language Contact*, 673–694. Oxford: Wiley-Blackwell.

Naylor, Paz Buenaventura. 1995. Subject, topic, and Tagalog syntax. In David Benett, Theodora Bynon & George Hewitt (eds.), *Subject, Voice and Ergativity*, 161–201. London: School of Oriental and African Studies.

Nerbonne, John. 2009. Data-Driven Dialectology. *Language and Linguistics Compass* 3(1). 175–198.

Nichols, Johanna. 1986. Head-marking and dependent-marking grammar. *Language* 62(1). 56–119.

Nichols, Johanna, Alena Witzlack-Makarevich & Balthasar Bickel. 2013. The AUTOTYP genealogy and geography database: 2013 release. http://www.spw.uzh.ch/autotyp/.

Pagès, Jérôme. 2004. Analyse factorielle de données mixtes. *Revue de Statistique Appliquée* 4. 93–111.

Pawley, Andrew & Harald Hammarström. 2018. The Trans New Guinea family. In Bill Palmer (ed.), The Languages and Linguistics of the New Guinea Area: a Comprehensive Guide, 21–196. *The World of Linguistics, Vol. 4*. Berlin: De Gruyter Mouton.

Plank, Frans. 1998. The co-variation of phonology with morphology and syntax: A hopeful history. *Linguistic Typology* 2. 195–230.

Polinsky, Maria. 2013. Applicative Constructions. In Matthew S. Dryer & Martin Haspelmath, (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/109.)

Poser, William J. 1995. Binary Comparison and the History of Hokan Comparative Studies. *International Journal of American Linguistics* 61(1). 135–144.

Ramat, Paolo. 1986. Is a Holistic Typology possible? *Folia Linguistica* 20 (1-2). 3–14.

Reid, Lawrence A. 2005. The current status of Austric: A review and evaluation of the lexical and morphosyntactic evidence. In Laurent Sagart, Roger Blench & Alicia Sanchez-Mazas (eds.), *The peopling of East Asia: putting together archaeology, linguistics and genetics*, 134–162. London: Routledge Curzon.

Sasse, Hans-Jürgen. 1984. The pragmatics of noun incorporation in eastern Cushitic languages. In Frans Plank (ed.), *Objects: toward a theory of grammatical relations*, 243-268. London: Academic Press.

Schlegel, August Wilhelm von. 1818. *Observations sur la langue et la litérature provençales*. Paris: Libraire grecque-latine-allemande

Schlegel, Karl Friedrich von. 1808. *Über die Sprache und Weisheit der Indier: Ein Beitrag zur Begründung der Alterthumskunde, nebst metrischen Übersetzungen indischer Gedichte*. Heidelberg: Mohr and Zimmer.

Schmidt, Wilhelm. 1906. Die Mon–Khmer-Völker, ein Bindeglied zwischen Völkern Zentralasiens und Austronesiens ('The Mon–Khmer Peoples, a link between the Peoples of Central Asia and Austronesia'). *Archiv für Anthropologie* 5. 59–109.

Schönig, Claus. 2003. Turko-Mongolic Relations. In Juha Janhunen (ed.), *The Mongolic Languages*, 403–419. London: Routledge.

Skirgård, Hedvig, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances* 9(16). 1–15. DOI: 10.1126/sciadv.adg6175

Song, Jae Jung. 2013. Nonperiphrastic Causative Constructions. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://wals.info/chapter/111.)

Whorf, Benjamin Lee. 1941. The Relation of Habitual Thought and Behavior to Language. In Leslie Spier, A. Irving Hallowell & Stanley S. Newman (eds.), *Language, Culture, and Personality: Essays in Memory of Edward Sapir*, 75–93. Menasha, Wisconsin: Sapir Memorial Publication Fund. Reprinted in John B. Carroll (ed.) 1956. Langu*age, Thought and Reality. Selected Writings of Benjamins Lee Whorf*. Cambridge, Mass: The MIT Press.

CONTACT

s.kalyan@uq.edu.au
mhdonohue@gmail.com.